QUALITY OF SERVICE AND MOBILITY MANAGEMENT IN IP-BASED RADIO ACCESS NETWORKS

LI FENG

(B.Eng.(Hons.), NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2003

ACKNOWLEDGEMENTS

Many thanks are given to my supervisors, Dr. Winston K.G. Seah and Dr Hoang M. Nguyen, for their guidance along the way, especially the discussions, critiques and advice during the course of paper and thesis writing.

Many thanks are due to lots of friends around ... Special thanks to Cheng Jing, Ng Keng Seng, Wu Wei, Xie Qunying and Yang Luqing, for their friendship, for the hardship and fun we had working together, for making the past two years a memorable experience!

The dissertation is dedicated to my parents and my sister. It is their love, support and encouragement that made everything I have possible!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS OR ABBREVIATIONS

3GPP	Third Generation Partnership Project.
3GPP2	Third Generation Partnership Project.
AAL2	ATM Adaptation Layer type 2.
AF	Assured Forwarding.
AP	Access Point.
ATM	Asychronous Transfer Mode.
BACK	Binding Acknowledge.
BS	Base Station.
BU	Binding Update.
CBR	Constant Bit Rate.
CDMA	Code Division Multiple Access.
CR-LDP	Constraint-based Routing Label Distribution Protocol.
DiffServ	Differentiated Services.
DSCP	DiffServ Code Point.
EF	Expedited Forwarding.
FA	Foreign Agent.
FEC	Forwarding Equivalence Class.
FIFO	First-In First-Out.
FP	Frame Protocol.
GERAN	GSM/EDGE Radio Access Network.
GFA	Gateway Foreign Agent.
GPRS	General Packet Radio System.
GSM	Global System for Mobile communications.
GTP	GPRS Tunnelling Protocol.
HA	Home Agent.
HMIPv6	Hierarchical Mobile IPv6.

IEEE	Institute of Electrical and Electronics Engineers.
IETF	Internet Engineering Task Force.
IMT-2000	International Mobile Telephony in 2000.
IntServ	Integrated Services.
IP	Internet Protocol.
IPHC	IP Header Compression.
LAN	Local Area Network.
LEMA	Label Edge Mobility Agent.
LER	Label Edge Router.
LMA	Local Mobility Agent.
LSP	Label Switched Path.
MAC	Media Access Control.
MAP	Mobility Anchor Point.
MIP	Mobile IP.
MPLS	MultiProtocol Label Switching.
MT	Mobile Terminal.
MWIF	Mobile Wireless Internet Forum.
OAM	Operation Adminstration and Management.
OSPF	Open Shortest Path First.
PHB	Per Hop Behavior.
QoS	Quality of Service.
RAN	Radio Access Network.
RIO	Random early drop with In/Out.
RIP	Routing Interior Protocol.
RNC	Radio Network Controller.
RNL	Radio Network Layer.
RSVP	Resource reSerVation Protocol.
RTP	Real-time Transport Protocol.

SONET	Synchronous Optical Network.
SPF	Shortest Path First.
TCP	Transmission Control Protocol.
TDMA	Time Division Multiple Access.
ToS	Type of Service.
TS	Technical Specification.
TSG	Technical Specification Group.
UE	User Equipment.
UMTS	Universal Mobile Telecommunications System.
UTRAN	UMTS Terrestrial Radio Access Network.
VoIP	Vocie over IP.
VPN	Virtual Private Network.
W-CDMA	Wideband Code Division Multiple Access.
WDM	Wavelength Division Multiplexing.
WRR	Weighted Round Robin.

SUMMARY

In third generation radio access networks (3G RANs), a transport technology is needed to interconnect the network elements such as base stations (BSs) and radio network controller (RNC). 3GPP release'99 specifies the use of ATM as the transport technology inside UTRAN (UMTS Terrestrial Radio Access Network). In recognition that using IP as the foundation for next generation mobile network makes strong economic and technical sense, there is a strong interest in IP-based transport in 3G RANs. IP-based solutions, however, face a number of challenges in QoS and mobility management in order to meet the stringent transport and control requirements of 3G RANs. Efficient transport techniques are needed to take best benefit of IP technologies to satisfy the diverse QoS requirements while efficiently utilizing network resources in the radio access networks; efficient micro-mobility techniques are needed, integrated with QoS support, to reduce latency, packet loss, and signaling overhead during handover.

A framework of QoS support in IP/MPLS-based radio access networks is proposed. Combining QoS routing with dynamic resource allocation, simulation results show that this can provide a suitable framework of QoS support in radio access networks. Moreover, a hierarchical MPLS-based micro-mobility scheme is proposed. By introducing one more level of hierarchy in hierarchical-based radio access network, optimal forwarding path is used for data packets upon handover and hence handover latency and packet loss are reduced. The interoperation of such micro-mobility scheme with Hierarchical Mobile IPv6 is described and simulation results show that this approach improves handover performance, in terms of handover latency and packet loss.

CHAPTER I

INTRODUCTION

In third generation radio access networks (3G RANs), a transport technology is needed to interconnect the network elements such as base stations (BSs) and radio network controller (RNC). Mobile Wireless Internet Forum (MWIF) has proposed an IP-based RAN (also known as OpenRAN) [1]. The IP-based RAN is a new version of future RAN architecture that is fully optimized to carry IP traffic. The division of functionality between network elements is fundamentally re-defined to suit the needs of IP traffic (e.g., radio functions such as macro-diversity and outer loop power control may be moved closer to BSs). A functional reference architecture for 3G IP-based RAN is presented in [2]:

- The *transport plane* provides basic routing functionality inside the RAN and gateway routing functionality towards other RAN and the core network. It also provides QoS enforcement in RAN and QoS admission control and mapping towards/from external networks.

- In the *control plane*, the Mobile Control Function is the central control function to manage dedicated logical radio resources, connection management, handover, and location management. A Micro-Mobility Anchor Function supports fast and loss less user plane relocation triggered by a mobile terminal (MT) mobility by providing a user plane anchor point and its management.

- In the *user plane*, IP packets carrying user payload arrive at the Micro Mobility Anchor that forwards the packets the appropriate User Radio Gateway hiding the RAN micro-mobility from the core network.

The main concept of IP-based RAN architecture is to separate the control plane, user plane, and transport plane so that the benefit of IP can be fully utilized, such as flexibility to allocate processing capacity for user traffic and control traffic in different locations; independently scale the control plane and the user plane by increasing/decreasing number of nodes required to handle the corresponding traffic. Most of simulations conducted by several companies consistently demonstrated that IP transport performance is equal or better than the ATM transport currently used in the RAN.

IP-based solutions, however, face a number of challenges in QoS and mobility management in order to meet the stringent transport and control requirements of 3G RANs. Issues concerning with enhancing the basic IP micro-mobility management protocols with scalable capabilities that reduce latency, packet loss, and signaling overhead during handover are inherent in wide-area mobility protocols. It is interesting to investigate and evaluate how micro-mobility can be provided in an efficient way with continuous QoS support.

1.1 QoS Support in RANs

QoS support in computer networks is essentially a resource allocation problem. In the context of radio access network, it is actually a traffic engineering problem! The major objectives of *Internet traffic engineering* are to enhance the performance of an operational network, at both the traffic and resource levels [3]. At the traffic level, both user plane and control plane traffic require the underlying transport bearers to support a variety of QoS requirements and traffic characteristics. At the resource level, network resources, in terms of link bandwidth, router buffers, are required to be utilized efficiently [4].

Indeed, in RANs efficient resource usage is a critical objective, since radio access network is precisely one of the most expensive part of the wireless network. The QoS problems in RANs can be thus described as *developing transport techniques that take best benefit of IP technologies to satisfy the diverse QoS requirements while efficiently*

utilizing network resources in the radio access networks.

1.2 Micro-Mobility Management in RANs

IP micromobility protocols [5] are designed for environments where mobile hosts change their point of attachment to the network so frequently that the base Mobile IP mechanism introduces significant network overhead in terms of increased delay, packet loss, and signaling. In the context of radio access network, it is also desired to develop efficient micro-mobility techniques, integrated with QoS support.

1.3 IP/MPLS-based Radio Access Networks

Multiprotocol Label Switching (MPLS) in conjunction with IP is known as IP/MPLS, which is substituted conventional IP address lookup and forwarding within a network by the faster operations of label lookup and switching. Label Switched Path (LSP) can be either signaled or engineered to provide QoS guarantees. Traffic engineered LSP can be provided with restoration paths for reliability, while LSP constructed using link state information are automatically re-configured whenever the state is refreshed. Moreover, the framework for signaling, traffic engineering, QoS, restoration, and virtual private networks (VPNs) is already available for MPLS networks and being actively deployed. The notable benefits of MPLS inspire our work on the use of MPLS to complement IP as a transport solution in the IP/MPLS-based RAN, with the objectives of support QoS and Micro-Mobility.

1.4 Contribution

Corresponding to the above mentioned research topic in RANs, the main contributions of this thesis are:

- A framework of QoS support in IP/MPLS-based radio access networks is proposed and presented in [6]. Combining QoS routing with dynamic resource allocation, simulation results show that this approach can provide a suitable framework for QoS support in IP/MPLS-based radio access networks.

- A hierarchical IP/MPLS-based micro-mobility scheme is proposed and presented in [7]. By introducing one more level of hierarchy in hierarchical-based radio access network, optimal forwarding path is used for data packets upon handover and hence handover latency and packet loss are reduced.

- The interoperation of the micro-mobility scheme [7] with Hierarchical Mobile IPv6 is presented and simulation results show that this approach improves handover performance, in terms of handover latency and packet loss.

1.5 Organization

The remainder of the thesis is organized as follows. Chapter 2 reviews relevant background information and related works. Chapter 3 presents an overall framework of QoS support in IP/MPLS-based radio access networks. Chapter 4 presents a hierarchical MPLS-based micro-mobility scheme with QoS support and its interoperation with hierarchical Mobile IPv6. Chapter 5 presents simulation results by ns-2, including performance comparison for two different Traffic Engineering approaches and three tunnel-based micro-mobility schemes. Finally Chapter 6 delivers some concluding remarks and the direction for future work.

CHAPTER II

BACKGROUND AND RELATED WORKS

2.1 Third Generation Mobile System

With over one billion mobile phone users estimated by the end of 2002, and packet-based multimedia services, including IP telephony, accounting for over 50 percent of all wireless traffic, it is natural to provide more capacity in the mobile network, and higher bandwidth in the radio link, radio access network (RAN), and core network. There is a momentum in the industry to evolve the current infrastructure, network services, and end-user applications toward an end-to-end IP solution capable of supporting quality of service (QoS) to meet the needs of the dominant data traffic. At the present time there are fundamentally three types of second-generation (2G) digital networks: Global System for Mobile Communications (GSM), time-division multiple access (TDMA), and code-division multiple access (CDMA). There are several 2.5G interim data transport standards, which are being pursued by many operators in their network implementations. Their decisions are based on many complex trade-offs, such as user demand, regulatory conditions (spectrum availability), cost (of equipment and spectrum license), backward compatibility, and their assessment of which will be the dominant 3G worldwide standard. A question worth answering is: what is 3G? It is mobile multimedia, personal services, the convergence of digitalization, mobility, the Internet, new technologies based on global standards, all of the above. The end user will be able to access the mobile Internet at the bandwidth (on demand) from hundreds of kilobits per second to about 2 Mb/s. From a business perspective it is the business opportunity of the 21st century [8].

There are several 2G to 3G evolution scenarios for the operators, and some would

be content with using 2.5G technologies to make their networks reach 3G characteristics and features. Although wideband CDMA (W-CDMA), also known as International Mobile Telecommunications in 2000 (IMT-2000) or Universal Mobile Telecommunications System (UMTS), has emerged as one of the leading standards, other flavors of 3G standards (e.g., CDMA-2000) are still being considered by some operators and countries.

There are certain differences in the approaches to specific aspects of the 3G network architecture, the most pronounced being those between the specifications introduced by the Third Generation Partnership Project (3GPP) and 3GPP2, the two leading wireless industry consortia.

The standard interfaces and components of a 3G UMTS network are outlined in TS 23.002 [9]. There are two land-based network segments: the UMTS radio access network (UTRAN) and the core network (CN). Together, they form the administrative domain of the mobile operator. The CN itself is further divided into the circuit- and packet-switched domains.

A mobile user's equipment (UE) communicates with multiple base stations, call node Bs in UMTS, over the wireless Uu interface [10]. They are referred as *access points* (APs) in accordance with the IETF terminology. The outgoing (uplink) user-level packets are segmented by the UE into radio network layer (RNL) frames, called *transport blocks*. These are carried over the radio frequency layer, using the wireless CDMA (W-CDMA) access and modulation techniques, to the APs within reach of the mobile. Each AP encapsulates a set of transport blocks into a single frame of the RNL framing protocol (FP) and forwards the frame to its radio network controller (RNC) over the Iub interface. The details of the sublayers of the RNL such as the packet data convergence protocol, radio link control, medium access control, and radio frequency layer are outlined in TS 25.401 [11].

When the multiple APs serving a mobile host (or UE) have different controlling

RNCs, one of the latter acts as the serving RNC for that host. The FP frames are exchanged between the controlling (drift) and serving RNCs over the Iur interface. The serving RNC of the host is responsible for frame selection among the multiple received copies of the same transport block, processing the other sublayers of the RNL, and finally reassembling the user-level packet. It also maintains the link layer state for the host, that is, it maps the host identity with the identities of the APs and the communication channels within each AP that currently serves that host. The transport network between the APs and the RNCs has been traditionally composed of point-to-point T1 lines.

2.2 OpenRAN

Cellular telephony networks depend on an extensive wired network between the core network and the radio transceivers that handle particular cells. This network, called a radio access network (RAN), provides functions that coordinate access to the radio link between multiple radio base stations and between mobile terminals.

Existing RAN architectures for cellular systems are based on a centralized radio network controller connected by point-to-point links with the radio base transceiver stations. According to [12], the existing architecture is subject to a single point of failure if the RNC fails, and is difficult to expand because adding an RNC is expensive. Also, although a network operator may have multiple radio link protocols available, most RAN architectures treat each protocol separately and require a separate RAN control protocol for each.

A new architecture for mobile wireless RANs, called the OpenRAN, is proposed by the Mobile Wireless Internet Forum (MWIF) IP in the RAN working group [1]. OpenRAN is based on a distributed processing model with a routed IP network as the underlying transport fabric, with the following architectural principles [2]:

- *Separation of transport and control*. The Internet's Minimal Network Intelligence paradigm [13] entails a clear separation of transport and applications. All data transport is based on end-to-end datagram delivery by means of protocols of the TCP/IP suit. Applying this principle to the RAN architecture implies making the logical separation of transport plane, user plane, and control plane. In this way, the control plane may be implemented by signaling servers on standard all-purpose host platforms; the user plane with its real-time radio processing requirements may be implemented on highly specialized hardware; the transport plane consists of a standard router network. As a further consequence, the separated planes scale separately depending on the operators' needs. Thus physical separation of transport, user, and control plane increases the deployment flexibility. As an additional benefit a cost reduction is expected if standard routing platforms are employed in the transport plane.

- *Distributed control architecture*. The client-server paradigm is the dominant communications model in the Internet. Applying this principle to the RAN architecture, the control plane may be distributed on several hosts. In this way, some of the control functions may be executed on standard server platforms while others may be placed on highly specialized hardware platforms. By doing so, the control functions scale independently of each other, and introducing new control elements by means of a server is simpler than extending a monolithic function block.

- *Open interfaces*. The open interface policy of the Internet allows virtually anybody to provide a service as long as she complies to the open interface standards. The open interface are the main means to utilize the creativity of the Internet community for new innovative services. Applied to the RAN architecture the open interface paradigm fosters a multi-vendor environment. The introduction of a new service into the network is simplified.

- *IETF standardized protocols.* Open interfaces are closely relate to standardized Internet protocols, provided by the Internet Engineering Task Force (IETF). Since these protocols are widely used on millions of hosts in the Internet an economies of scales effect may be expected from applying them in the RAN as well instead of developing new protocols for a small number of special purpose hosts.

2.3 Internet QoS

QoS is a set of technologies that enables network administrators to manage the effects of congestion on application traffic by using network resources optimally, rather than by continually adding capacity [14].

Applications generate traffic at varying rates and generally require that the network be able to carry traffic at the rate at which they generate it. In addition, applications are more or less tolerant of traffic delays in the network and variation in traffic delay. Certain applications can tolerate some degree of traffic loss, while others cannot. If infinite network resources were available, all application traffic could be carried at the applications required rate, with zero latency and zero packet loss. However, network resources are not infinite. As a result, there are parts of the network in which resources are unable to meet demand. Networks are built by concatenating network devices such as switches and routers. They forward traffic among themselves using interfaces. If the rate at which traffic arrives at an interface exceeds the rate at which that interface can forward traffic to the next device, congestion occurs. Thus, the capacity of an interface to forward traffic is a fundamental network resource. QoS mechanisms work by allotting this resource preferentially to certain traffic over other traffic. In order to do so, it is first necessary to identify different classes of traffic. Traffic arriving at network devices is separated into distinct flows via the process of packet classification. Traffic from each flow is then directed to a corresponding queue on the forwarding interface. Queues on each interface are serviced according to some algorithm. The queue servicing algorithm determines the rate at which traffic from each queue is forwarded, thereby determining

the resources allotted to each queue and to the corresponding flows. Thus in order to provide network QoS, it is necessary to provision or configure the following in network devices:

- Classification information by which devices separate traffic into flows
- Queues and queue servicing algorithms that handle traffic from the separate flows

The above can be jointly referred to as *traffic handling mechanisms* [14]. Traffic handling mechanisms in isolation, are not useful. They must be provisioned or configured across many devices in a coordinated manner that provides useful end-to-end services across a network. To provide useful services, therefore, requires both *traffic handling mechanisms* and *provisioning and configuration mechanisms*. The provisioning and configuration mechanisms coordinate traffic handling mechanisms subject to *policies* that are devised by network administrators.

Work on QoS-enabled IP networks has led to two distinct approaches: the Integrated Services architecture (IntServ) [15] and its accompanying signaling protocol, RSVP [16], and the Differentiated Services architecture (DiffServ) [17].

2.3.1 Integrated Services and RSVP

The integrated services (IntServ) architecture defined a set of extensions to the traditional best effort model of the Internet with the goal of allowing end-to-end QOS to be provided to applications. One of the key components of the architecture is a set of service definitions; the current set of services consists of the controlled load and guaranteed services. The architecture assumes that some explicit setup mechanism is used to convey information to routers so that they can provide requested services to flows that require them. While RSVP is the most widely known example of such a setup mechanism, the IntServ architecture is designed to accommodate other mechanisms.

RSVP [16] is a signaling protocol that can be used by hosts to request resource reservations through a network. RSVP can be considered a mechanism for configuring

traffic handling mechanisms in network devices. IntServ assumes that network devices support traffic handling mechanisms, which guarantee service to each traffic flow in strict isolation from other traffic flows. It also assumes services that offer specific quantities of resources. In 1997 the RSVP working group of the IETF was busy putting the finishing touches on the RSVP protocol design. At the same time, the IntServ working group was defining the services that could be expected by applications in response to RSVP signaling. Another working group, the Integrated Services over Specific Link Layers (issll), was defining the underlying traffic handling mechanisms that would offer QoS support on different media. As this work was unfolding, the media was busily hyping RSVP as a panacea - the magic cure that would bring an end to all network woes.

As is often the case with over-hyped technologies, RSVP and IntServ failed to deliver on the promises. There are a number of reasons for this:

- RSVP was supposed to be signaled by hosts, but only experimental versions of the protocol were available and only on certain UNIX platforms.
- There was a perception that RSVP and IntServ had to be implemented on every network device and that it was not scalable.
- There were no policy mechanisms to govern, in a secure manner, which traffic flows were granted privileged access to network resources.
- RSVP and IntServ focused on protecting multimedia applications and not on the non-multimedia mission-critical applications that were (and still are) considered more important by network administrators.

IntServ identifies three main categories of services that can be provided to users. *Guaranteed services* [18] provide users with an assured amount of bandwidth, firm end-to-end delay bounds, and no queuing loss for flows. Controlled load [19] services assure that the users will get service that is as close as possible to the one received by a

best-effort service in a lightly loaded network. Best effort services are characterized by absence of a QoS specification and the network delivers the best possible quality.

2.3.2 Differentiated Services

Differentiated services (DiffServ) was born against the backdrop of RSVP's fall from grace. It promised to overcome the scalability concerns of RSVP. DiffServ was greeted with great enthusiasm, by both the IETF and by the media, hungry for a new panacea. DiffServ is a traffic handling mechanism. It defines a field in packets IP headers, called the DiffServ codepoint (DSCP) [20]. Hosts or routers sending traffic into a DiffServ network mark each transmitted packet with a DSCP value. Routers within the DiffServ network use the DSCP to classify packets and apply specific queuing behavior (known as per-hop behavior or PHB) based on the results of the classification. Traffic from many flows having similar QoS requirements is marked with the same DSCP, thus aggregating the flows to a common queue or scheduling behavior.

The distinguishing feature of DiffServ is its scalability. To understand DiffServ's inherent scalability, it is important to contrast aggregate traffic handling mechanisms versus per-conversation traffic handling mechanisms. The traffic handling mechanisms envisioned in RSVP/IntServ networks are per-conversation mechanisms. These treat each traffic flow between each instance of a sending and receiving application, in isolation. Aggregate mechanisms, such as DiffServ, group many traffic flows into a single aggregate class. Another aggregate traffic handling mechanism is the use of 802.1p prioritization in IEEE 802 networks. Per-conversation traffic handling mechanisms rely on per-conversation classifiers. These typically use the IP source and destination addresses and ports to uniquely identify conversations. Aggregate traffic handling mechanisms typically rely on some mark in a packet that aggregates it into a queue shared by other packets with the same mark. In the examples of DiffServ and 802.1p, these marks are DSCPs or 802.1p tags, respectively. Before the packet is submitted to the aggregate traffic handling mechanism, it must be marked with the appropriate tag.

Aggregate traffic handling mechanisms require significantly less state and processing power in network nodes than their per-conversation counterparts. This benefit is most significant in large networks that are required to handle large numbers of conversations. The compromise of aggregate traffic handling is that the QoS enjoyed by each conversation is dependent on the behavior of the other conversations with which it is aggregated.

DiffServ is the preferred technology for large-scale IP QoS deployments today, such as service provider backbone networks. DiffServ achieves scalability through performing complex QoS functions such as classification, marking, and conditioning operations at the edges of the network. Traffic is classified and then marked using the DSCP into a limited number of traffic aggregates or classes. Within the core of the network, scheduling and queuing control mechanisms are applied to the traffic classes based upon the DS field marking; all traffic conditioning and dropping is handled intelligently at the network layer using IP DiffServ quality of service mechanisms. DiffServ is not prescriptive in defining the scheduling and queuing control algorithms that should be implemented at each hop, but rather, uses a level of abstraction in defining the externally observable forwarding behaviors, termed PHBs, that can be applied to traffic at each hop. Currently, three PHBs are defined:

- *The expedited forwarding (EF) PHB.* The EF PHB [21] is used to support traffic with low loss, low delay, low jitter, assured bandwidth requirements, such as VoIP.

- *The assured forwarding (AF) PHB.* The AF PHB [22] is used to support data traffic with assured bandwidth requirements.

- *The default PHB.* This PHB [20] represents the default forwarding behavior. Packets, which are not identified as belonging to another class, belong to this aggregate.

2.3.3 Roles of IntServ, RSVP and DiffServ

IntServ, RSVP and DiffServ are viewed as complementary technologies in the pursuit of end-to-end QoS [14]. Together, these mechanisms can facilitate deployment of applications such as IP-telephony, video- on-demand, and various non-multimedia mission-critical applications. IntServ enables hosts to request per-flow, quantifiable resources, along end-to-end data paths and to obtain feedback regarding admissibility of these requests. DiffServ enables scalability across large networks.

2.4 Internet Traffic Engineering and QoS Routing

Internet traffic engineering [3] is defined as that aspects of Internet network engineering dealing with the issue of performance evaluation and performance optimization of operational IP networks. Traffic Engineering encompasses the application of technology and scientific principles to the measurement, characterization, modeling, and control of Internet traffic.

Enhancing the performance of an operational network, at both the traffic and resource levels, are major objectives of Internet traffic engineering. This is accomplished by addressing traffic oriented performance requirements, while utilizing network resources economically and reliably. Traffic oriented performance measures include delay, delay variation, packet loss, and throughput.

QoS-based routing [23], or a routing mechanism under which paths for flows are determined based on some knowledge of resource availability in the network as well as the QoS requirement of flows, has been recognized as a missing piece in the evolution of QoS-based service offerings in the Internet.

2.4.1 Best-Effort and QoS-Based Routing

Routing deployed in today's Internet is focused on connectivity and typically supports only one type of datagram service called "best effort" [24]. Current Internet routing protocols, e.g. OSPF, RIP, use "shortest path routing", i.e. routing that is optimized for

a single arbitrary metric, administrative weight or hop count. These routing protocols are also "opportunistic", using the current shortest path or route to a destination. Alternate paths with acceptable but non-optimal cost cannot be used to route traffic (shortest path routing protocols do allow a router to alternate among several equal cost paths to a destination). QoS-based routing extends the current routing paradigm in three basic ways [23].

- to support traffic using integrated-services class of services, multiple paths between node pairs will have to be calculated. Some of these new classes of service will require the distribution of additional routing metrics, e.g. delay, and available bandwidth. If any of these metrics change frequently, routing updates can become more frequent thereby consuming network bandwidth and router CPU cycles.

- today's opportunistic routing will shift traffic from one path to another as soon as a "better" path is found. The traffic will be shifted even if the existing path can meet the service requirements of the existing traffic. If routing calculation is tied to frequently changing consumable resources (e.g. available bandwidth) this change will happen more often and can introduce routing oscillations as traffic shifts back and forth between alternate paths. Furthermore, frequently changing routes can increase the variation in the delay and jitter experienced by the end users.

- as mentioned earlier, today's optimal path routing algorithms do not support alternate routing. If the best existing path cannot admit a new flow, the associated traffic cannot be forwarded even if an adequate alternate path exists.

2.4.2 QoS-Based Routing and Resource Reservation

It is important to understand the difference between QoS-based routing and resource reservation. While resource reservation protocols such as RSVP [16] provide a method

for requesting and reserving network resources, they do not provide a mechanism for determining a network path that has adequate resources to accommodate the requested QoS. Conversely, QoS-based routing allows the determination of a path that has a good chance of accommodating the requested QoS, but it does not include a mechanism to reserve the required resources.

Consequently, QoS-based routing is usually used in conjunction with some form of resource reservation or resource allocation mechanism. Simple forms of QoS-based routing have been used in the past for Type of Service (ToS) routing [25]. In the case of OSPF, a different shortest-path tree can be computed for each of the 8 TOS values in the IP header [26]. Such mechanisms can be used to select specially provisioned paths but do not completely assure that resources are not overbooked along the path. As long as strict resource management and control are not needed, mechanisms such as TOS-based routing are useful for separating whole classes of traffic over multiple routes. Such mechanisms might work well with the emerging Differential Services efforts [17].

Combining a resource reservation protocol with QoS-based routing allows fine control over the route and resources at the cost of additional state and setup time. For example, a protocol such as RSVP may be used to trigger QoS-based routing calculations to meet the needs of a specific flow.

2.5 MultiProtocol Label Switching

Multiprotocol Label Switching (MPLS) is a promising effort to provide the kind of traffic management and connection-oriented *Quality of Service* (QoS) support found in *Asynchronous Transfer Mode* (ATM) networks, to speed up the IP packet-forwarding process, and to retain the flexibility of an IP-based networking approach.

The roots of MPLS go back to numerous efforts in the mid-1990s to combine IP and ATM technologies. The first such effort to reach the marketplace was IP switching,

developed by Ipsilon. To compete with this offering, numerous other companies announced their own products, notably Cisco Systems (Tag Switching), IBM (aggregate route-based IP switching), and Cascade (IP Navigator). The goal of all these products was to improve the throughput and delay performance of IP, and all took the same basic approach: Use a standard routing protocol such as Open Shortest Path First (OSPF) to define paths between endpoints; assign packets to these paths as they enter the network; and use ATM switches to move packets along the paths. When these products came out, ATM switches were much faster than IP routers, and the intent was to improve performance by pushing as much of the traffic as possible down to the ATM level and using ATM switching hardware.

In response to these proprietary initiatives, the Internet Engineering Task Force (IETF) set up the MPLS working group in 1997 to develop a common, standardized approach. The working group issued its first set of Proposed Standards in 2001. Meanwhile, however, the market did not stand still. The late 1990s saw the introduction of many routers that are as fast as ATM switches, eliminating the need to provide both ATM and IP technology in the same network.

Nevertheless, MPLS has a strong role to play. MPLS reduces the amount of per-packet processing required at each router in an IP-based network, enhancing router performance even more. More significantly, MPLS provides significant new capabilities in four areas that have ensured its popularity: QoS support, traffic engineering, Virtual Private Networks (VPNs), and multiprotocol support.

- *Connection-Oriented QoS Support*. A connectionless network, such as in IP-based internetwork, cannot provide truly firm QoS commitments. A Differentiated Service (DiffServ) framework [17] works in only a general way and upon aggregates of traffic from numerous sources. An Integrated Services (IntServ) framework [15], using the *Resource Reservation Protocol* (RSVP), has some of the flavor of a connection-oriented approach, but is nevertheless limited in terms

of its flexibility and scalability. For services such as voice and video that require a network with high predictability, the DiffServ and IntServ approaches, by themselves, may prove inadequate on a heavily loaded network. By contrast, a connection-oriented network has powerful traffic management and QoS capabilities. MPLS imposes a connection-oriented framework on IP-based internet and thus provides the foundation for sophisticated and reliable QoS traffic contracts.

- *Traffic Engineering*. MPLS makes it easy to commit network resources in such a way as to balance the load in the face of a given demand and to commit to differential levels of support to meet various user traffic requirements. The ability to dynamically define routes, plan resource commitments on the basis of known demand, and optimize network utilization is referred to as traffic engineering. With the basic IP mechanism, there is a primitive form of automated traffic engineering. Specifically, routing protocols such as OSPF enable routers to dynamically change the route to a given destination on a packet-by-packet basis to try to balance load. But such dynamic routing reacts in a very simple manner to congestion and does not provide a way to support QoS. All traffic between two endpoints follows the same route, which may be changed when congestion occurs. MPLS, on the other hand, is aware of not just individual packets, but flows of packets in which each flow has certain QoS requirements and a predictable traffic demand. With MPLS, it is possible to set up routes on the basis of these individual flows, with two different flows between the same endpoints perhaps following different routers. Further, when congestion threatens, MPLS paths can be rerouted intelligently. That is, instead of simply changing the route on a packet-by-packet basis, with MPLS, the routes are changed on a flow-by-flow basis, taking advantage of the known traffic demands of each flow. Effective use of traffic engineering can substantially increase usable network capacity.

- *VPN Support.* MPLS provides an efficient mechanism for supporting VPNs. With

a VPN, the traffic of a given enterprise or group passes transparently through an internet in a way that effectively segregates that traffic from other packets on the internet, proving performance guarantees and security.

- *Multiprotocol Support.* MPLS, which can be used on many networking technologies, is an enhancement to the way a connectionless IP-based internet is operated, requiring an upgrade to IP routers to support the MPLS features. Routers can coexist with ordinary IP routers, facilitating the introduction of evolution to MPLS schemes. MPLS is also designed to work in ATM and Frame Relay networks. Again, MPLS-enabled ATM switches and MPLS-enabled Frame Relay switches can be configured to coexist with ordinary switches. Furthermore, MPLS can be used in a pure IP-based internet, a pure ATM network, a pure Frame Relay network, or an internet that includes two or even all three technologies. This universal nature of MPLS should appeal to users who currently have mixed network technologies and seek ways to optimize resources and expand QoS support.

2.5.1 Overview

As a packet of a connectionless network layer protocol travels from one router to the next, each router makes an independent forwarding decision for that packet, i.e., each router analyzes the packet's header, and each router runs a network layer routing algorithm. Each router independently chooses a next hop for the packet, based on its analysis of the packet's header and the results of running the routing algorithm [27].

Packet headers contain considerably more information than is needed simply to choose the next hop. Choosing the next hop can therefore be thought of as the composition of two functions. The first function partitions the entire set of possible packets into a set of "Forwarding Equivalence Classes (FECs)". The second maps each FEC to a next hop. Insofor as the forwarding decision is concerned, different packets which get mapped into the same FEC are indistinguishable. All packets which belong to a particular FEC and which travel from a particular node will follow the same path.

In conventional IP forwarding, a particular router will typically consider two packets to be in the same FEC if there is some address prefix X in that router's routing tables such that X is the "longest match" for each packet's destination address. As the packet traverses the network, each hop in turn reexamines the packet and assigns it to a FEC.

In MPLS, the assignment of a particular packet to a particular FEC is done just once, as the packet enters the network. The FEC to which the packet is assigned is encoded as a short fixed length value known as a "label". When a packet is forwarded to its next hop, the label is sent along with it, i.e., the packets are "labeled" before they are forwarded. At subsequent hops, there is no further analysis of the packet's network layer header. Rather, the label is used as an index into a table which specifies the next hop, and a new label. The old label is replaced with the new label, and the packet is forwarded to its next hop.

In the MPLS forwarding paradigm, once a packet is assigned to a FEC, no further header analysis is done by subsequent routers; all forwarding is driven by the labels. This has a number of advantages over conventional network layer forwarding:

- MPLS forwarding can be done by switches which are capable of doing label lookup and replacement, but are either not capable of analyzing the network layer headers, or are not capable of analyzing the network layer headers at adequate speed.

- Since a packet is assigned to a FEC when it enters the network, the ingress router may use, in determining the assignment, any information it has about the packet, even if that information cannot be gleaned from the network layer header. For example, packets arriving on different ports may be assigned to different FECs. Conventional forwarding, on the other hand, can only consider information which travels with the packet in the packet header.

- A packet that enters the network at a particular router can be labeled differently than the same packet entering the network at a different router, and as a result

forwarding decisions that depend on the ingress router can be easily made. This cannot be done with conventional forwarding, since the identity of a packet's ingress router does not travel with the packet.

- The considerations that determine how a packet is assigned to a FEC can become ever more and more complicated, without any impact at all on the routers that merely forward labeled packets.

- Sometimes it is desirable to force a packet to follow a particular route which is explicitly chosen at or before the time the packet enters the network, rather than being chosen by the normal dynamic routing algorithm as the packet travels through the network. This may be done as a matter of policy, or to support traffic engineering. In conventional forwarding, this requires the packet to carry an encoding of its route along with it ("source routing"). In MPLS, a label can be used to represent the route, so that the identity of the explicit route need not be carried with the packet.

Some routers analyze a packet's network layer header not merely to choose the packet's next hop, but also to determine a packet's "precedence" or "class of service". They may then apply different discard thresholds or scheduling disciplines to different packets. MPLS allows (but does not require) the precedence or class of service to be fully or partially inferred from the label. In this case, one may say that the label represents the combination of a FEC and a precedence or class of service.

2.5.2 MPLS and Traffic Engineering

MPLS is strategically significant for Traffic Engineering because it can potentially provide most of the functionality available from the overlay model, in an integrated manner, and at a lower cost than the currently competing alternatives. Equally importantly, MPLS offers the possibility to automate aspects of the Traffic Engineering function.

This last consideration requires further investigation and is beyond the scope of this manuscript.

According to [28], a traffic trunk is an aggregation of traffic flows of the same class which are placed inside a Label Switched Path. Essentially, a traffic trunk is an abstract representation of traffic to which specific characteristics can be associated. It is useful to view traffic trunks as objects that can be routed; that is, the path through which a traffic trunk traverses can be changed. In this respect, traffic trunks are similar to virtual circuits in ATM and Frame Relay networks. It is important, however, to emphasize that there is a fundamental distinction between a traffic trunk and the path, and indeed the LSP, through which it traverses. An LSP is a specification of the label switched path through which the traffic traverses. In practice, the terms LSP and traffic trunk are often used synonymously.

The attractiveness of MPLS for Traffic Engineering can be attributed to the following factors [29]:

- explicit label switched paths which are not constrained by the destination based forwarding paradigm can be easily created through manual administrative action or through automated action by the underlying protocols,
- LSPs can potentially be efficiently maintained,
- traffic trunks can be instantiated and mapped onto LSPs,
- a set of attributes can be associated with traffic trunks which modulate their behavioral characteristics,
- a set of attributes can be associated with resources which constrain the placement of LSPs and traffic trunks across them,
- MPLS allows for both traffic aggregation and disaggregation whereas classical destination only based IP forwarding permits only aggregation,

- it is relatively easy to integrate a "constraint-based routing" framework with MPLS,
- a good implementation of MPLS can offer significantly lower overhead than competing alternatives for Traffic Engineering.

Additionally, through explicit label switched paths, MPLS permits a quasi-circuit switching capability to be superimposed on the current Internet routing model. Many of the existing proposals for Traffic Engineering over MPLS focus only on the potential to create explicit LSPs. Although this capability is fundamental for Traffic Engineering, it is not really sufficient. Additional augmentations are required to foster the actualization of policies leading to performance optimization of large operational networks.

2.6 Mobility Management

Mobility management has widely been recognized as one of the most important and challenging problems for a seamless access to wireless networks and mobile services. It is the fundamental technology used to automatically support mobile terminals enjoying their services while simultaneously roaming freely without the disruption of communications. Two main aspects need to be considered in mobility management, i.e., location management (e.g. addressing, location registration and update, tracking and paging, etc.) and handover management (e.g. handover trigger and initiation, connection routing, smoothing, etc.).

Classification of mobility protocol can be achieved regarding many of their characteristics. It can be assumed that they share a common goal of overcoming the location dependent nature of IP addresses of Internet hosts by developing mechanisms for translation of addresses and efficient distribution of packets to and from any location both for static and highly Mobile Hosts [30]. Concerning their scope, the current mobility protocols can be classified into two main categories: global or macro-mobility protocols and micro- or regional-mobility protocols.

2.6.1 Macro-mobility Management

The mobile IP [31] protocol is the current standard for supporting macroscopic mobility in IP networks, i.e. host mobility across IP domains while maintaining transport level connections. It is transparent for applications and transport protocols, which work equal with fixed or mobile hosts. It can be scaled to provide mobility across the Internet. And it allows nodes using mobile IP inter-operate with nodes using the standard IP. There are two versions of mobile IP: mobile IPv4 [31] and mobile IPv6 [32]. Each one addresses a particular version of IP.

2.6.2 MIPv4 vs. MIPv6

Mobile IP (MIP) supports mobility of IP hosts by allowing them to make use of (at least) two IP addresses: a home address that represents the fixed address of the node and a care-of-address (CoA) that changes with the IP subnet the mobile node is currently attached to.

An entity that maps a home address to the corresponding currently valid CoA. In MIPv4 [31], these mappings are exclusively handled by “home agents” (HA). A corresponding node (CN) that wants to send packets to a mobile node (MN) will send the packets to the MN’s home address. In the MN’s home network these packets will be “intercepted” by the home agent and tunneled, e.g., by IP-in-Ip encapsulation [33], either directly to the MN or to a foreign agent to which the MN has a direct link. In MIPv6 [32], home agents no longer exclusively deal with the address mapping, but each CN can have its own “binding cache” where home address plus CoA pairs are stored. This enables “route optimization” compared to the triangle routing via the HA in MIPv4: a CN is able to send packets directly to a MN when the CN has a recent entry for the MN in its corresponding binding cache. When a CN sends a packet directly to a MN, it does not encapsulate the packet as the HA does, but makes use of the IPv6 Routing Header Option. When the CN does not have a binding cache entry for the MN, it sends the packet to the MN’s home address. The MN’s home agent will then forward the

packet. The MN when receiving an encapsulated packet will inform the corresponding CN about the current CoA.

In order to keep the home address to CoA mappings up-to-date, a mobile node has to signal corresponding changes to its corresponding nodes and/or home agent when performing a handover to another IP subnet. Since in MIPv6 both, CN and HA, maintain binding caches, a common message format called “binding updates” is used to inform CN and HA about changes in the point of attachment. Additionally, MIPv6 allows a MN to send a binding update to a MIPv6 agent in the IP subnet previously visited by the MN. Then, packets sent by CNs that have not yet learned the MN’s new CoA will be tunneled from the previously visited subnet to the current point to attachment. Binding updates (BU) can be acknowledged by BU ACKs. In contrast to MIPv4, where signaling is done in extension headers that can also be piggybacked on “regular” packets. To acquire a CoA in MIPv6, a mobile node can build on IPv6 stateless and stateful auto-configuration methods. The stateless auto-configuration mechanism is not available in IPv4.

2.6.3 Micro-mobility Management

For the support of regional-mobility within one domain or one site, the mobile IP solution was found non-optimal. Firstly, it generates significant signalling traffic in the core network even for local movement. Secondly, it creates a considerable delay in the diffusion of mobile hosts localization updates. And finally, it causes long interruptions and packet losses during handovers. Therefore a new protocol providing the management of micro-mobility seems to be necessary.

Existing proposals for micromobility can be broadly classified into two types: routing-based and tunnel-based schemes [5]. *Routing-based schemes* aim to exploit the robustness of conventional IP forwarding. A distributed mobile host location database is created and maintained within the network domain. The database consists of individual flat mobile-specific address lookup tables and is maintained by all the mobility agents

within the domain. These schemes are exemplified by the Cellular IP [34] and Hawaii [35] protocols, which differ from each other in the functionality of the nodes and the construction methods of the lookup tables. In one form or another, the *tunnel-based schemes* apply the concepts of registration and encapsulation in a local or hierarchical fashion, thus creating a flexible concatenation of (possibly several) local tunnels. In the context of MIPv4, the Mobile IP regional registration proposal [36] falls into this category. Hierarchical Mobile IPv6 [37] plays a similar role in IPv6 networks. An early example of a tunnel-based scheme is provided by GTP-based mobility management in GPRS and UMTS.

2.6.4 Handover Management

Handover refers in general to support for terminal mobility wherever the mobile host changes its point of attachment to the network [30]. More specifically, the access network may provide particular capabilities to minimize the interruption to sessions in progress. In wireless mobile networks different handover scenarios might occur. A Layer 2 handover happens if the network layer is not involved in the handover; intra-access network handover when the new point of attachment is in the same access network; inter-access network handover when the new access router is in a different access network. Horizontal or vertical handover are said to happen if the old and the new access router use the same or different wireless interface (technology) respectively. Others handover types can be defined according to different phases of the handover. Three phases are distinguished in the handover:

- *Initiation phase* The objective of this phase is to recognize the need for a handover and subsequently initiate it. The handover can be required by the mobile host or by the network. Generally, it is initiated when the radio link quality between a mobile host and its access router is degraded. However, it can also be initiated for network management and maintenance reasons. For example, in case of overload some mobile hosts may be moved from an access router to another one.

- *Decision phase* In this phase, measurements on neighboring radio transmitters and eventual network policy information are first collected. Then the best target access router is identified taken into account the measurement and information report. The execution phase is finally triggered to perform the corresponding handover. According to whether the mobile host or the network handles these operations, four handover types are differentiated: mobile or network controlled handover if the mobile host or the network initiates and decides a handover. Network-assisted handover when the network collects information that can be used in handover decisions and mobile-assisted handover when information and measurements from the mobile host are used to decide the execution of a handover.

- *Execution phase* In the execution phase, the mobile host has been detached from the old access router and attached to the new one. The order of attach and detach events is not fixed. During a soft handover the mobile host communicates simultaneously with the old and the new access router whereas in a hard handover it is not able to do it. Handover may imply re-routing of connections through the fixed network and an address negotiation for the mobile host like the acquisition of a new care-of-address and the registration procedure in mobile IP. In planned handover, contrary to unplanned handover, some signalling messages can be sent before the mobile host is connected to the new access router, e.g. building a temporary tunnel from the old access router to the new access router. If the handover is initiated via the currently serving access router, it is a backward handover, else it is a forward handover.

2.7 Related Works

There have been some works done on QoS and micro-mobility in IP/MPLS-based RANs [38, 39, 40, 41], but none of them are complete or fully validated.

In [38], a DiffServ-based approach has been proposed. All traffic are transported between RNC and BS in a single, statically provided stream, which is mapped onto an Expedited Forwarding (EF) traffic class. The simulation results only show that DiffServ-based IP-transport provides a good alternative to ATM-transport in the UTRAN.

A MPLS-based transport scheme has been proposed in [39], which incorporates constraint-based routing and DiffServ to provide transport bearers that can support bandwidth provisioning and a variety of QoS requirements in the RAN. The basic idea is to establish and manage label switching paths (LSPs) for interconnecting base stations (BSs) and a radio network controller (RNC) in UTRAN. Constraint-based routing label distribution protocol (CR-LDP) is used to set up LSPs based on bandwidth constraint. Differentiated QoS is provided: 1) A single LSP carries multiple class of traffic; 2) Multiple LSPs, each LSP carries one class of traffic. However, [39] did not show any simulation results, also this proposed scheme only considers the transport for data traffic, not for the more critical signaling traffic. Moreover, issues of mobility and integration with QoS were not considered in this scheme.

In [40], a framework associated with signaling for intra-domain micro-mobility using label switched path re-direction in a traffic engineered network. An enhanced label edge router (LER) called the label edge mobility agent (LEMA) is introduced to support chained LSP-redirection. The scheme is scalable and suitable for fast handover, transient packet loss associated with local movement, QoS support, and gradual evolution. However, the algorithms for choosing the LEMAs for a particular mobile host is still not clear, and mobile host has the burden of keeping extra information, such as the previous LEMA registered, the previous BS attached. Moreover, IP/MPLS technology is terminated at the access routers causes a limitation of this approach into OpenRAN.

[41] proposes a scheme to integrate the Mobile IP and MPLS protocols, which improves the scalability of the Mobile IP data forwarding process by leveraging on the features of MPLS, and remove the need for IP-in-IP tunneling from HA to FA. However, the scheme is still targeted at traditional Mobile IPv4, and also suffer the concern

of non-applicability for micro-mobility, as the scope of Mobile IP is more for global mobility.

CHAPTER III

IP-RAN TRAFFIC ENGINEERING

This chapter presents a framework of QoS support for IP/MPLS-based radio access networks (presented in [6]). DiffServ is chosen as an IP QoS differentiation model with MPLS (MultiProtocol Lable Switching) as the underlying forwarding scheme: both BSs and RNC are DiffServ edge routers and MPLS Label Edge Routers (LERs), whereas the other interior routers are both DiffServ core routers and MPLS Label Switching Routers (LSRs). Two levels of service differentiation are defined, namely MPLS-level and DiffServ-level. Two approaches of QoS support, namely DiffServ-based and MPLS-based, or IP-TE and MPLS-TE, are studied through simulation to investigate the possible problems for QoS support in IP-based RANs, as well as evaluate the performance in terms of throughput and delay (results are presented in Chapter 5, Section 5.2).

3.1 QoS in Radio Access Networks

The fundamental concept of UMTS/IMT-2000 is the separation of the access functionality from the core network functionality. The RAN provides an access platform for mobile terminals (MTs) to all core networks and network services. It hides all radio-access-technology-dependent and mobility functions from the core network. The two types of RANs currently in the scope of 3GPP are UTRAN and GSM/EDGE radio access network (GERAN), based on WCDMA and EDGE radio access technologies, respectively.

In 3G RANs, a transport technology is needed to interconnect the network elements such as base stations (BSs) and radio network controllers (RNCs). The diverse QoS requirements of the applications themselves (e.g., real-time or non-real-time) combined with the requirements imposed by advanced radio control functions (e.g., soft handover

and power control in CDMA systems) require that the transport technologies provide differentiated QoS to multiple classes of traffic. The transport bearers need to support a variety of QoS requirements (delay, jitter, packet loss, etc.) and traffic characteristics (streaming, bursty, etc.).

In particular, the WCDMA radio control functions and real-time applications impose rather stringent delay requirements on the UTRAN transport network:

- For real-time traffic, the tight end-to-end delay of the applications along with many other components in the delay budget impose rather stringent UTRAN transport delay requirements. It is specified as less than 7ms in the current 3GPP specification [8].

- For non-real-time traffic, the UTRAN transport delay is governed by the radio functions, in particular outer-loop power control and soft-handoff control. For outer-loop power control to function properly, the round trip delay is preferably less than 50 ms, corresponding to a one-way delay of 25 ms [8]. This requires the transport delay to be less than 10 ms (this value is for future study). For soft-handoff control, the two branches for macrodiversity combining must be synchronized, and larger delay will increase the complexity of maintaining the synchronization between the soft-handoff branches.

The jitter requirement for UTRAN transport is not specified as a specific value but in general should be less than 10 percent of the transport delay. The loss ratio for UTRAN transport should be at least one order less than that of the air interface, so for voice traffic it should be less than 1e-4, and for data traffic less than 1e-7. It should be noticed that the figures given here are exemplary rather than exact numbers. The stringent requirements on delay, jitter, and loss ratio indicate that UTRAN transport is a “real-time mission-critical” application of the transport network. It should be given very high priority and firm commitment of resources in the transport network.

3.1.1 ATM-based Transport Solutions

Among various packet networking technologies, ATM currently has relatively mature schemes to support QoS. In the first UMTS releases, ATM/AAL2 (ATM Adaptation Layer 2) [42] is chosen as the transport technology in the UTRAN. In order to meet the stringent QoS requirements (e.g., delay and packet loss) in the UTRAN transport layer, special attention must be given to network dimensioning, traffic management, and resource management. In [43] a number of issues related to the performance and design of the ATM/AAL2 transport in the UTRAN are addressed. The effect of the stringent delay requirement on the bandwidth requirement of ATM/AAL2 transport is studied. The simulation results suggest that the delay requirement for ATM/AAL2 transport should not be too stringent in order to avoid poor bandwidth utilization caused by packet-scale congestion. A traffic management scheme using Common Part Sublayer (CPS) packet shaping is proposed to deal with the burstiness of constant bit rate (CBR) traffic caused by the periodic nature of the medium access control (MAC) layer of the UMTS radio interface. The simulation results show that CPS packet shaping will significantly reduce the bandwidth requirement. The statistical multiplexing gain from ATM/AAL2 transport over TDM transport is evaluated and found to be significant. The overall results confirm that with careful network dimensioning, traffic management, and resource management, ATM/AAL2 technology is capable of meeting the stringent QoS requirements in the WCDMA UTRAN.

3.1.2 IP-based Transport Solutions

While ATM/AAL2 has relatively mature schemes to support QoS, there is a strong interest in alternative technology such as IP-based transport in 3G RANs. In this scheme the network elements of 3G RAN (e.g., base stations and radio network controllers) are interconnected via an IP network, as shown in Figure 3.1.

There are several motivations for the use of IP transport in the RAN: IP QoS management is approaching maturity; IP as a network layer protocol is carefully designed

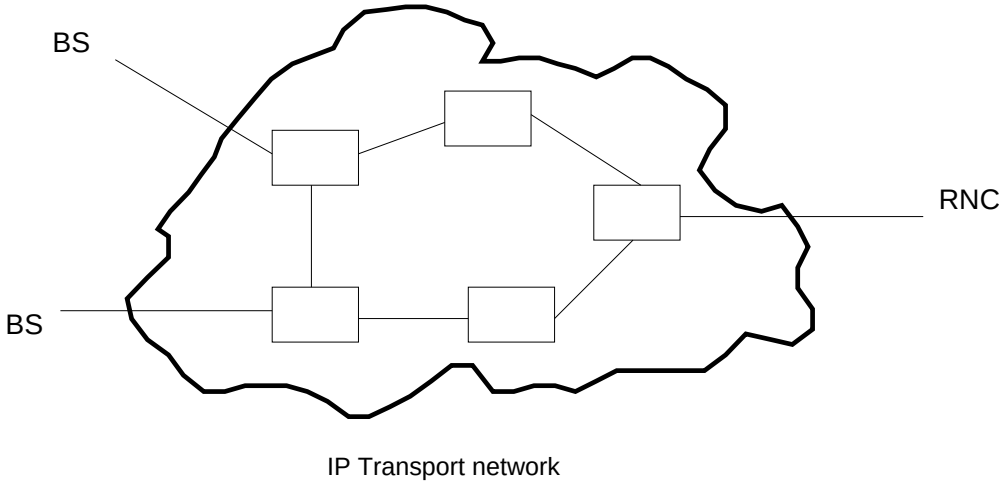


Figure 3.1: IP transport in the RAN

to be independent of link/physical layers, so it allows a wide selection of lower-layer technologies, including options of IP over synchronous optical network (SONET) or IP over wavelength-division multiplexing (WDM); IP is quickly becoming the basis for packetization of voice, data, signaling, and operation, administration, and management (OAM) in the networking world. Another important fact is that the 3G core network is IP-based; therefore, an IP-based RAN will allow consistent backbone infrastructure, operational efficiency, and industry standard OAM.

IP-based transport solutions face a number of challenges in order to meet the stringent transport requirements of the 3G RANs, especially the WCDMA UTRAN. In terms of QoS this translates to tight end-to-end control of delay and jitter, and almost zero packet loss ratio. Current IP networks were designed for delay-insensitive data applications. IP-based transport solutions must be enhanced to provide QoS support including delay, jitter, and loss. It should also support real-time signaling transport, as well as reliability and security. Transport efficiency for a qualified IP solution is another important issue. Since the RTP/UDP/IP header could be larger (about 60 bytes) than that of ATM (5 bytes), it is a concern that the IP header overhead is much higher than the ATM overhead for transporting voice. This must be clarified and addressed properly.

Underlying technologies to enable an IP-based RAN are evolving into maturity at a

fast pace. Internet routers have become faster (with latency less than that of time switching technology, i.e., < 125 ms), more flexible (supporting priority queues essential for service differentiation), and more robust (commercially available fault-tolerant reliable routers). Progress in IETF standardization is pushing IP transport to a technically viable solution for the 3G RAN. These include: RTP (transport protocol for real-time packet streaming), SigTran (real-time signaling transport over IP), IntServ (Integrated Services Architecture for Guaranteed QoS), RSVP (mechanism for reserving dedicated bandwidth/router resources for QoS management), DiffServ (architecture for scalable service differentiation), MPLS (label switching with traffic engineering capabilities), and IPHC (IP header compression for reducing overhead).

IP-based transport solutions for 3G RANs are being studied in 3GPP and Mobile Wireless Internet Forum (MWIF). The "IP Transport in UTRAN" is being specified in Technical Specification Group (TSG) RAN working group 3 in 3GPP. The "IP in the RAN" technical group in MWIF is focusing on using IP as a transport technology for various RANs, including 3GPP RAN, 3GPP2 RAN, and others [44].

3.1.3 QoS support in RAN: IP-RAN Traffic Engineering

QoS support in computer networks is essentially a resource allocation problem. In the context of radio access network, it is actually a traffic engineering problem! The major objectives of Internet traffic engineering are to enhance the performance of an operational network, at both the traffic and resource levels [3]. At the traffic level, both user plane and control plane traffic require the underlying transport bearers to support a variety of QoS requirements and traffic characteristics. At the resource level, network resources, in terms of link bandwidth, router buffers, are required to be utilized efficiently [4].

3.2 Framework of QoS Support for IP/MPLS-based Radio Access Networks

Described in [6], a framework of QoS support for IP/MPLS-based radio access networks is proposed. Combining QoS routing with dynamic resource allocation, simulation results (presented in Chapter 5, Section 5.2) show that it can provide a suitable framework of QoS support for IP/MPLS-based radio access networks. There are two main components for the framework, namely, traffic performance and resource utilization.

3.2.1 Traffic Performance with Dynamic Resource Allocation in DiffServ-enabled MPLS Networks

Because of its simplicity and scalability, DiffServ [17] is chosen for QoS differentiation model integrated as the underlying transport bearers for both user and control plane traffic in IP/MPLS-based radio access networks. Two-levels of differentiation are defined to address the different traffic QoS requirements with the combination of static and dynamic LSP configuration.

- At MPLS-level, separate LSPs are established for control plane traffic and user plane traffic to cater the different transport requirements between the two types of traffics.
- At DiffServ-level, MPLS service classes are defined according to the 3-bit EXP field of MPLS packet header for mapping of DiffServ Per Hop Behaviors (PHBs). Control plane traffic and real-time user plane traffic are mapped to DiffServ Expedited Forwarding (EF) PHB; non-real-time user plane traffic is mapped to DiffServ Assured Forwarding (AF) PHB.

Control plane traffic is essential for the proper operation of the overall network. Inherently for 3G-like wireless network, control plane traffic between BSs and RNCs

has very tight delay constraints, and is considered as the highest priority. The LSP for control plane traffic are static, i.e. pre-established with static resource reservation.

User plane traffic is the serving target of a proper operating networks, and different transport requirements are needed for different kinds of traffic, e.g., real-time traffic must have tight delay constraints while non-real-time traffic does not but certain level throughput is required. As user plane traffic is driven by user demands and are dynamics inherently. The LSPs for user plane traffic are dynamically established on demand, i.e., resources are dynamically allocated.

3.2.2 Efficient Resource Utilization with QoS Routing

QoS routing [23], referred as a routing mechanism under which paths for flows are determined based on some knowledge of resource availability in the network as well as the QoS requirement of flows. This can be applied to solve the unnecessary network congestion problem caused by traditional IP Shortest-Path-First (SPF). A lot of works have been done or ongoing to investigate the performance of different QoS-routing algorithms for Internet traffic engineering [24]. The application of QoS routing in MPLS is somehow termed as constraint-based routing, but essentially they are trying to address the same problem.

The framework proposed considers using explicitly routed paths with the constraint of bandwidth, i.e., every hop on the path has sufficient bandwidth. Different routing mechanisms are used for different types of LSPs. For dynamic LSPs, i.e. the LSPs for user plane traffic, QoS routing algorithms, or explicit routing algorithms, are used to find the bandwidth-constrained route. While for static LSPs, i.e. the LSPs for control plane traffic, there is no need to use QoS routing, as those LSPs are pre-established so that the traditional routing algorithm, like SPF, can be used to find the route. For the exact QoS routing algorithms, or explicit routing algorithms, to be used, there are many options [45], such as *shortest-widest path algorithm* [24], *widest-shortest path algorithm* [46], etc.

After the route/path is found, MPLS signalling protocol is used to set up the route with resource (bandwidth) reservation. Currently there are two options for the signaling protocol: CR-LDP (Constraint-based LSP Setup using LDP) [47] and RSVP-TE (Extensions to RSVP for LSP Tunnels) [48]. Both can be used for MPLS label distribution and resource reservation.

CHAPTER IV

IP-RAN MICRO-MOBILITY

This chapter presents a hierarchical micro-mobility scheme integrated with QoS support for IP/MPLS-based 3G radio access network (presented in [7]) and its interoperation with Hierarchical Mobile IPv6 (HMIPv6). DiffServ is chosen for QoS differentiation model integrated with MPLS as the underlying forwarding scheme for micro-mobility. An enhanced label edge router (LER) called Local Mobility Agent (LMA) is introduced to set up two-stage label switched paths (LSPs) between radio network controller (RNC) and base station (BS) for reducing handover latency caused by local mobility. The method of locating LMA in the network is described and the associated signaling procedures, such as registration, LSP setup, resource management are proposed accordingly. Simulation results for HMIPv6 interoperation are presented in Chapter 5, Section 5.3.

4.1 Integration of MPLS-based QoS and Micro-Mobility

4.1.1 Traffic Performance with Resource Allocation in DiffServ-enabled MPLS Networks

In [6], a framework of QoS support in IP/MPLS-based radio access networks has been proposed. DiffServ is chosen for QoS differentiation model in this framework and integrated with MPLS as the underlying transport bearers for both user and control plane traffic in IP/MPLS-based radio access networks. Two-levels of differentiation are defined to address the different traffic QoS requirements with the combination of static and dynamic LSP configuration.

At MPLS-level, separate LSPs are established for control plane traffic and user plane traffic to cater the different transport requirements between the two types of traffics. The

LSPs for control plane traffic are pre-established with static resource reservation, and the LSPs for user plane traffic are dynamically established on demand, i.e., resource are dynamically allocated.

At DiffServ-level, MPLS service classes are defined according to the 3-bit EXP field of MPLS packet header for mapping of DiffServ PHBs. Control plane traffic and real-time user plane traffic are mapped to DiffServ Expedited Forwarding (EF) PHB; non-real-time user plane traffic is mapped to DiffServ Assured Forwarding (AF).

4.1.2 Hierarchical Micro-Mobility with QoS

4.1.2.1 Network Architecture

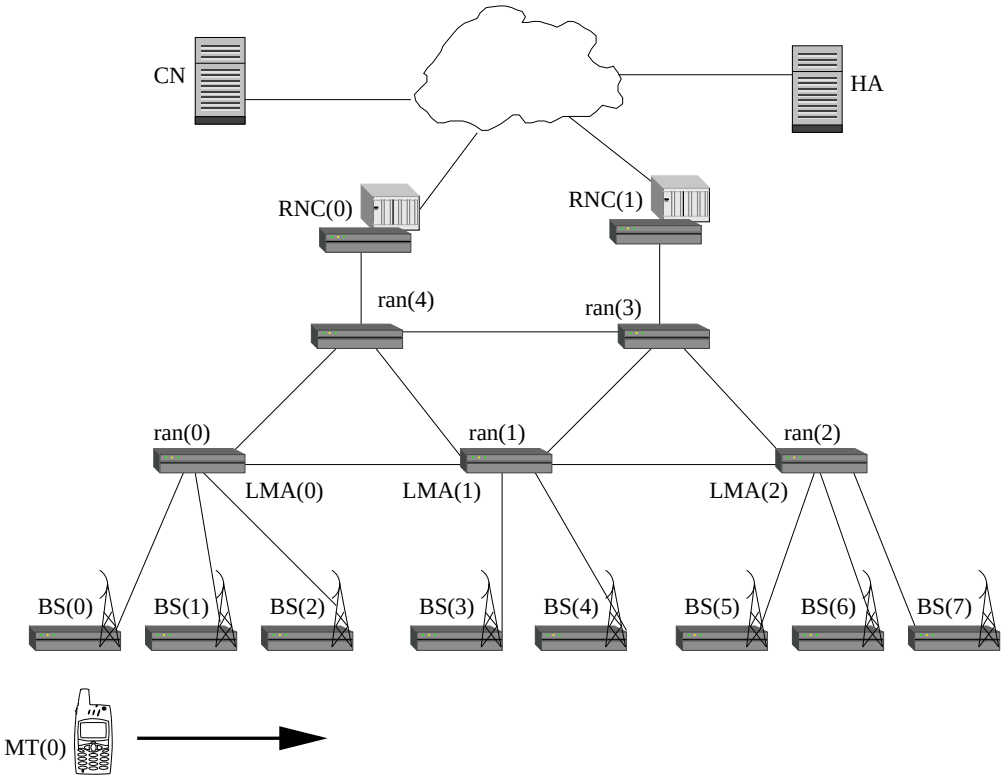


Figure 4.1: Hierarchical Radio Access Network

A hierarchical structured network is proposed for integration of micro-mobility and QoS in IP/MPLS-based radio access networks. Figure 4.1 shows a simplified version of this proposed network structure. There are three types of network entities for the integration of micro-mobility and QoS, represent different levels of hierarchy in the

access network.

RNC is the gateway of the network domain, standing for the highest-level of hierarchy in the network domain. A BS can directly communicate with mobile terminals (MTs), and is the lowest level of hierarchy. Both RNC and BS are label edge routers (LERs). A middle-level of hierarchy is an enhanced LER, called Local Mobility Agent (LMA), has been proposed to provide micro-mobility.

The basic idea of this micro-mobility scheme is hierarchical-based local registration with the support from LMA to bridge RNC and BS. Once the position of LMA is located in the network by the method described in Section 4.1.2.2, two stage static LSPs with resource reservation are pre-established between RNC and LMA, LMA and BS, respectively. Through these two stage LSPs, handover latency caused by local mobility can be reduced and continuous QoS requirements can be maintained.

4.1.2.2 Method of Locating LMA

The purpose of locating LMA is to find the nearest common router to BSs so that the common part of LSP can be reused during handover caused by local movement. Consequently, the time cost on re-establishing LSP can be reduced, and hence handover latency is reduced.

According to the QoS support framework [6], static LSPs with resource reservation are pre-established between every pair of RNCs and BSs for the transport of control plane traffics. The static LSPs, or the explicit routes, can be recorded at BS or RNC. By comparing the explicit routes from BSs to RNC, an nearest common router for each pair of BSs can be found and stored at BSs. This common router is selected as an LMA for those BSs passing through, or the higher level of hierarchy for those BSs crossing at the LMA.

Alternatively, during the period of pre-establishing static LSPs between RNCs and BSs, BS can send registration message hop-by-hop to all RNCs, and the first *ran* router receives more than one registration message from BSs can be selected as the LMA for

those BSs. For example in Figure 4.1, BS(0), BS(1), BS(2) send registration messages to RNC(0) and RNC(1). The first common router *ran*(0) will receive all three messages and is regarded as the LMA for BS(0), BS(1), and BS(2). Similarly, *ran*(1) is selected for BS(3) and BS(4), and *ran*(2) for BS(5), BS(6), BS(7).

4.2 Signaling Framework

With the assumption that LMAs have been properly located in the network, mobile terminal can be attached to one BS, and then LMA, RNC. It is the design objective of this scheme that network support for mobility should be maximized and burden for mobile terminal should be minimized. For example, there is no need for mobile terminal to record any information about the attaching BS, or the LMA consequently attached. This can reduce the burden on the mobile terminal, also better suits the integration of this micro-mobility scheme with macro-mobility protocol like Mobile IP. Whenever in registration or handover process, mobile terminal only need to send information to attaching BS about its home agent address, and its home address.

4.2.1 Registration and LSP Setup

Figure 4.2 shows the process of registration. Upon completion of the link layer attachment, an MT receives an advertisement message (ADS_BS) from a BS and sends registration message (REG_REQ_MT) to that BS with its home IP address, home agent IP address. The BS then registers with its LMA via the pre-established LSP (Section 4.1.2.2). When the corresponding LMA receives the request message from the lower-level BS, it adds its record about that MT and attaching BS. After that, LMA send registration message (REG_REQ_LMA) to its corresponding RNC via the pre-established LSP (Section 4.1.2.2) with the MT's home IP address, MT's home agent IP address, LMA's IP address. When the RNC receives the request message from the lower-level LMA, it adds its record about that MT and attaching BS. After that, RNC sends mobility binding message to MT's home agent with MT's home IP address, RNC's IP

address. When Binding ACK is received from MT's home agent at the RNC, this RNC will send registration reply message (REG_REPLY_RNC) back to the requesting LMA in its record, and consequently registration reply message (REG_REPLY_LMA) back to BS, and registration reply message (REG_REPLY_BS) back to mobile terminal.

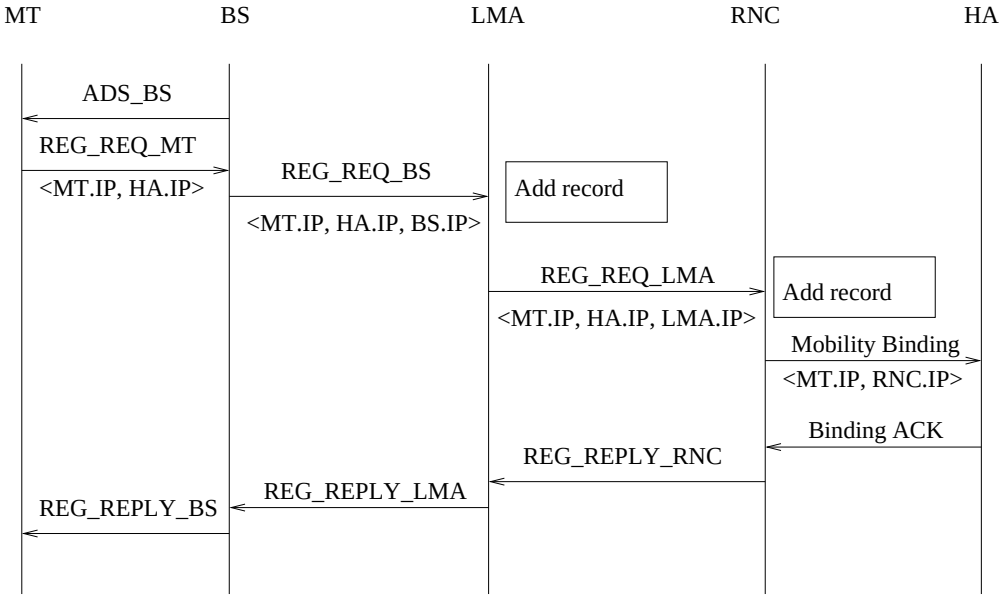


Figure 4.2: Registration process in Hierarchical Radio Access Network

Table 4.1: Registration: Record at LMA

MT	LMA	BS	LSP	PHB	Reservation
MT(0)	-	BS(0)	-	-	-

Table 4.2: Registration: Record at RNC

MT	LMA	BS	LSP	PHB	Reservation
MT(0)	LMA(0)	-	-	-	-

Table 4.1 and 4.2 show an example record kept in LMA(0) and RNC(0) when mobile terminal MT(0) in Figure 4.1 registers with BS(0), and then LMA(0), and then RNC(0). Noting at this time, this is neither data LSP established for this mobile terminal, nor any resource reservation record.

When a connection is initiated from/to the MT, new LSP tunnels, or data LSPs, are set up with a bandwidth reservation between RNC to LMA, as well as between LMA to BS. LSP setup can be initiated by either RNC or BS, depending on the direction of traffic flow and the signaling protocol being used. Considering downlink traffic and the use of CR-LDP, RNC will initiate the setup by sending LDP Request message downlink to LMA; then LMA sends LDP Mapping message back to RNC. In case the LSP is already established from RNC to LMA, only resource reservation is added accordingly but no new LSP is established: the same LSP is reused for data packets to the MT. Similarly, the LSP from LMA to BS is also established with resource reservation. The records at LMA(0) and RNC(0) are both updated to reflect the data LSP and any resource reservation, including DiffServ PHB requested (Table 4.3 and 4.4).

Table 4.3: Data LSP: Record at LMA

MT	LMA	BS	LSP	PHB	Reservation
MT(0)	-	BS(0)	LMA(0) → BS(0)	EF	100Kbps

Table 4.4: Data LSP: Record at RNC

MT	LMA	BS	LSP	PHB	Reservation
MT(0)	LMA(0)	-	RNC(0) → LMA(0)	EF	100Kbps

4.2.2 Handover and Partial LSP Re-direction

There are three types of handover in the considered hierarchical radio access network: Intra-BS, Intra-LMA, Inter-LMA. Intra-BS is basically link-layer handover, while Figure 4.3 shows the process of Intra-LMA handover, where both new BS and old BS are under the same LMA. Figure 4.4 shows the process of Inter-LMA handover, where new BS and old BS are under different LMAs.

In the case of Intra-LMA (Figure 4.3), when the LMA receives the registration message from BS, it checks its record and finds that the MT is already recorded with a

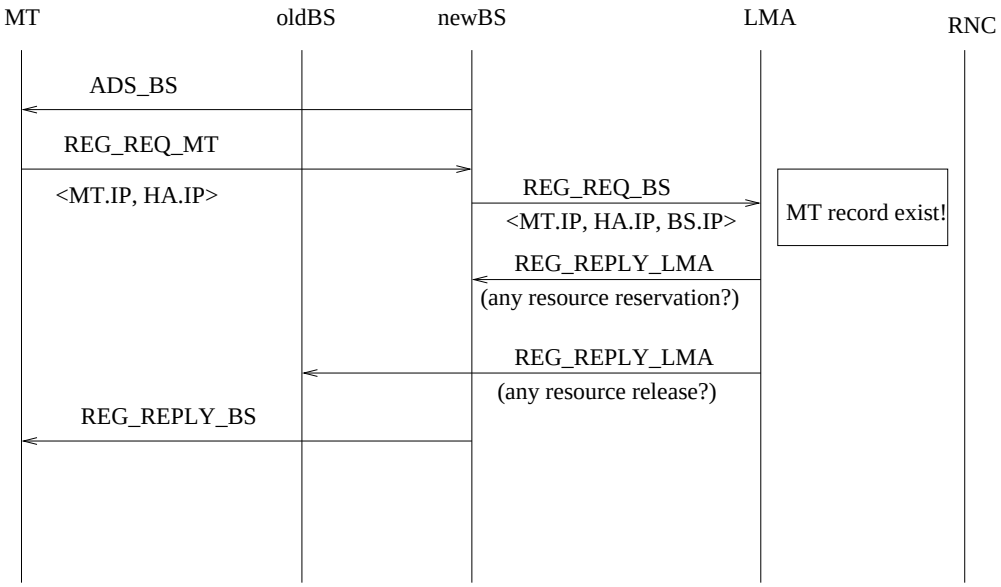


Figure 4.3: Intra-LMA Handover in Hierarchical Radio Access Network

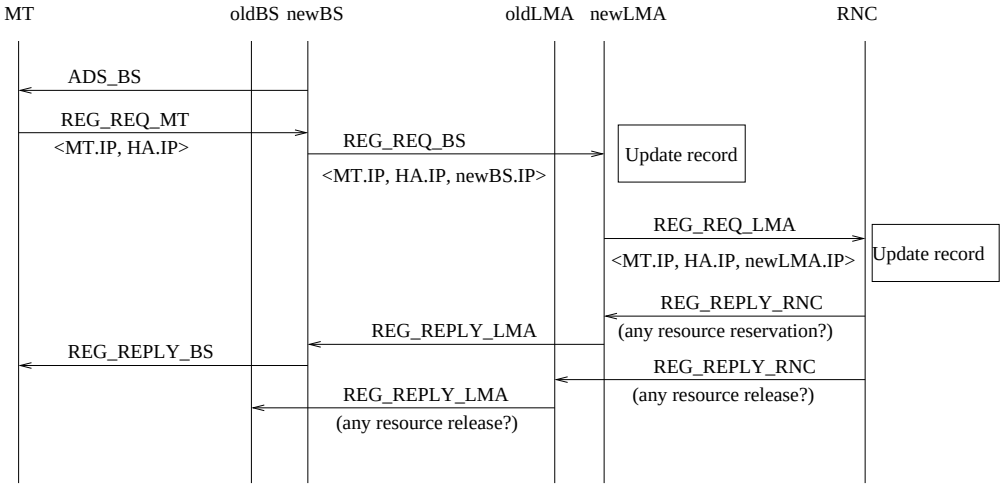


Figure 4.4: Inter-LMA Handover in Hierarchical Radio Access Network

different BS. The LMA then updates its record of that MT, and send a registration reply message to the new BS. In the meantime, data LSP to the new BS is established with resource reservation. Then the LMA send registration reply to the old BS for releasing any resources being reserved. Note that in this case, no message is sent to RNC to take advantage of common path revealed by the LMA. In this way, only partial LSP is re-directed and handover latency is reduced. Table 4.5 shows the record in LMA(0) when MT(0) in Figure 4.1 moves from BS(0) to BS(1).

In the case of Inter-LMA (Figure 4.4), when the new LMA receives the registration

Table 4.5: Intra-LMA Handover: Record at LMA

MT	LMA	BS	LSP	PHB	Reservation
MT(0)	-	BS(1)	LMA(0) → BS(1)	EF	100Kbps

message from the new BS, it finds there is no record for that MT. The LMA then adds its record of that MT, and send registration request message (REG_REG_LMA) to the corresponding RNC. When the RNC receives this request message, it checks its record and finds that a different LMA is attached. The RNC updates its record. For the RNC which has recorded data-LSP for that MT, it sends back reply message to the new LMA to setup new data-LSP with resource reservation. In the meantime, the resource reserved for that MT along the old data-LSP is released and the old data-LSP may be released if no other MT is attaching to that LSP. The old LMA is released with any resource data LSP to the new BS is established with resource reservation. Note that in this case, no message is sent to MT's home agent as only local movement is considered in this paper. When the new LMA receives the reply message from RNC, it sends reply message to the new BS with resource reservation. Table 4.6 and 4.7 show the record in LMA(0) and RNC(0) when MT(0) in Figure 4.1 moves from BS(2) to BS(3).

Table 4.6: Inter-LMA Handover: Record at LMA

MT	LMA	BS	LSP	PHB	Reservation
MT(0)	-	BS(3)	LMA(1) → BS(3)	EF	100Kbps

Table 4.7: Inter-LMA Handover: Record at RNC

MT	LMA	BS	LSP	PHB	Reservation
MT(0)	LMA(1)	-	RNC(0) → LMA(1)	EF	100Kbps

4.3 Interoperation with Hierarchical MIPv6

While routing-based schemes [5] avoid the tunneling overhead, they face difficulties in scaling, since for each mobile the forwarding table entries have to be replicated in all nodes on the uplink path, as opposed to selected nodes as in tunnel-based schemes. This also means that gradual deployment of routing-based mobility support can be difficult. Furthermore, the root (gateway) node of routing-based schemes constitutes a single point of failure. On the contrary, in the tunneling-based schemes it is possible to designate multiple Gateway Foreign Agents (GFAs) or Mobility Anchor Point (MAPs) within the micromobility domain, thus achieving higher robustness. All these factors, along with the ability to employ lightweight tunnels, explain why hierarchical tunnels seem to emerge as a preferred solution for supporting micromobility in all-IP wireless networks [10].

As a tunnel-based micromobility approach, Hierarchical Mobile IPv6 (HMIPv6) [37] is designed to minimize the amount of signaling to corresponding(s) and to the home agent by allowing the mobile host to locally register in a domain, and different hierarchical tunnels, e.g. IP tunnel (IP encapsulation [33]) and MPLS tunnel (LSP tunnel [27]) can be used. While Section 4.2 describes a generic singaling framework for the operation of MPLS-based micro-mobility scheme, it is later realized that the introduction of LMA for multiple-stages LSPs, called **MPLS-Tunnel with LMA**, would be a good lightweight yet fast tunneling approach for interoperation with HMIPv6.

With the assumptions that LMAs have been properly located in the network, **MPLS-tunnel with LMA** can inter-operate with HMIPv6 with minimal changes/enhancements to MIPv6. Such enhancement, and the operation of LMA/RNC for such interoperation are described as followed. Simulation results for the performance and comparison of different tunnels for HMIPv6 interoperation are presented in Chapter 5, Section 5.3.

4.3.1 Enhancements to MIPv6

For the inter-operation of **MPLS-Tunnel with LMA** with HMIPv6, an *LMA Bit* and an *LMA Field* are proposed to be added in MIPv6 Binding Update (BU) message. *LMA Bit*, a single bit, is used to indicate whether BU has traversed any previous LMA; *LMA Field*, an IP address field, is used to record the address of previous LMA which BU has traversed. *LMA Bit* is set by an LMA with its address recorded at *LMA Field*. Note that no additional messages are added and the basic MIPv6 location registration messages, BU and Binding Acknowledge (BACK) are reused.

Besides the additions to BU message, a record of *nextler* will be stored at LMA and RNC for each mobile hosts that have sent BU uplink with a *timer*. *nextler* refers to the previous LER (Label Edge Router), and is used for LMA/RNC to direct/redirect data packets of a mobile host to the right LSP. *timer* is set for each record of *nextler* to avoid possible scalability problem caused by the increased number of mobile hosts. The record for one mobile host will be refreshed upon receiving new BU from that mobile host and the *timer* is reset; when the *timer* expires, the record will be deleted.

4.3.2 Local Mobility Agent Operation

LMA "snoops" incoming BU message and is able to identify occurrence of handover by checking its record of *nextler* for the mobile host who sent the BU. Figure 4.5 shows the operation of Local Mobility Agent.

LMA checks whether *LMA Bit* is set. If no, BS address where the packet is coming from is added/updated as *nextler*; otherwise the address in *LMA Field* is added/updated as *nextler*. After that, *LMA Bit* is set and *LMA Field* is set with its own address. Then the record *timer* is reset and the BU message is forwarded uplink towards RNC/Gateway.

Based on this simple operation, LMA is able to identify when handover happens: when the record of *nextler*, or the previous LER is changed, corresponding to a handover, the new LER will be updated as *nextler* at the LMA. The LMA then can redirect the LSP before BU reaches RNC, i.e. the nearest crossover point could be identified so

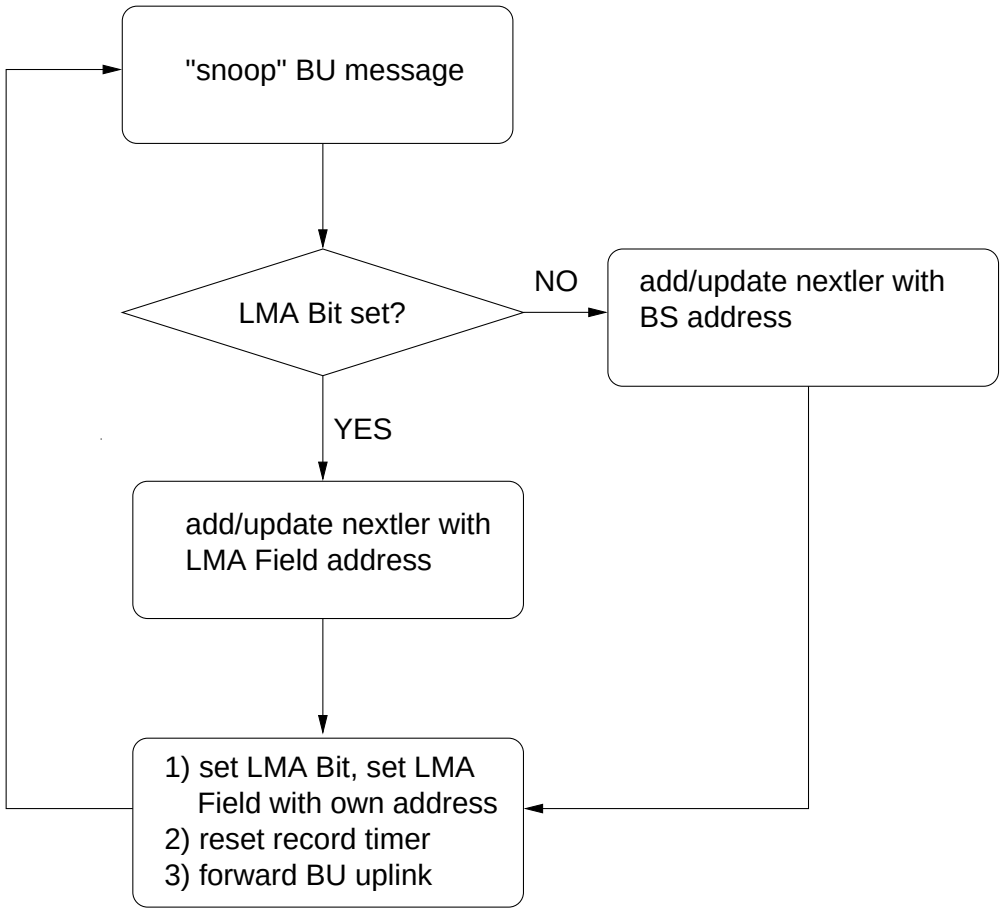


Figure 4.5: Local Mobility Agent Operation

that handover latency can be reduced.

4.3.3 RNC Operation

Similar to LMA, RNC will check whether *LMA Bit* is set. If no, BS address where the packet is coming from is added/updated as *nextler*; otherwise the address in *LMA Field* is added/updated as *nextler*. Then the record timer is reset and the handling of the BU message is up to HMIPv6 [37], i.e. forwarding BU to Home Agent and/or Correspondent Node, or sending BACK back for local mobility.

For the case that *LMA Bit* is not set, i.e, the nearest crossover point is at RNC, this corresponds the basic tunneling in HMIPv6 and no handover latency can be reduced.

CHAPTER V

SIMULATION RESULTS

This chapter presents simulation results for IP-RAN traffic engineering (Chapter 3) and IP-RAN micro-mobility (Chapter 4, Section 4.3) interoperated with HMIPv6. In Section 5.2, two different DiffServ-based QoS support approaches, namely **IP-TE** and **MPLS-TE**, are compared in terms of total network throughput, average packet delay, and per-DSCP average delay. In Section 5.3, three different tunnel-based micromobility approaches, namely **IP-Tunnel**, **MPLS-Tunnel** and **MPLS-Tunnel with LMA**, are compared in terms of handover latency, packet loss ratio and average delay

5.1 Simulation Tools

The Network Simulator ns-2 (version 2.1b9) [49] is used as the simulation tool, and a few contributed modules are added and integrated for the simulation, including MPLS module (MNS) [50, 51], MIPv6 module (MobiWan) [52], QoS Routing module [53]. Necessary modification/porting has been done for some modules for integrated simulation. For example, the MPLS module is extended to be used with hierarchical addressing which is necessary for wired-cum-wireless Mobile IP simulation; MIPv6 module is ported to ns-2.1b9 (originally ns-2.1b6 only, and not working with MPLS) and integrated with the MPLS module.

5.2 Simulation on IP-RAN Traffic Engineering

A simulation model was constructed in ns-2 to evaluate the performance of a system based on the framework presented in the Chapter 3. The main concern to use MPLS technology is about setup delay of LSP, i.e, how fast the label distribution process can be done without affecting the ongoing communication session. With the use of MPLS for

Traffic Engineering (**MPLS-TE**) in IP/MPLS-based DiffServ network, a comparison is made with **IP-TE** (IP Traffic Engineering) in Diffserv network: same QoS Routing algorithm is applied and IP source routing is used to route the packets along the explicit route found by QoS routing algorithm.

5.2.1 Link sharing and resource allocation in DiffServ Network

Table 5.1 shows the basic link sharing in the assumed DiffServ Network. Priority scheduling is used with EF PHB and the rest PHBs, and Weighted Round Robin (WRR) is used to share the rest of the bandwidth among AF PHB and BE (Best-Effort, or Default) PHB. FIFO (First-In First-Out) buffer management is used for EF and BE PHB, while RIO (Random Early Drop with In/Out) is used for AF PHB. Resource, especially link bandwidth, are dynamically allocated through admission control at the edge of the DiffServ network. QoS routing is used to find the route with maximum available bandwidth.

Table 5.1: DiffServ Link sharing and bandwidth allocation

PHB	Priority	Queuing	Scheduling
EF	Level 0	FIFO	PRIORITY between Level 0 and 1
AF	Level 1	RIO	WRR within the same priority
BE	Level 1	FIFO	WRR within the same priority

5.2.2 RAN Modeling and Assumptions

The RAN model for the simulation is shown in Figure 5.1. Four BS nodes are used to model the Base Station in IP-RAN, with the functionality of IP/MPLS forwarding. They are the “gateway” for mobile terminal to send/receive data. Two RNC nodes are used to model the Radio Network Controller in IP-RAN, also with the functionality of IP/MPLS forwarding. They are distributed network-wide, i.e., a BS can send traffic to any RNC, depends the available processing power in the RNC intended. Five Core

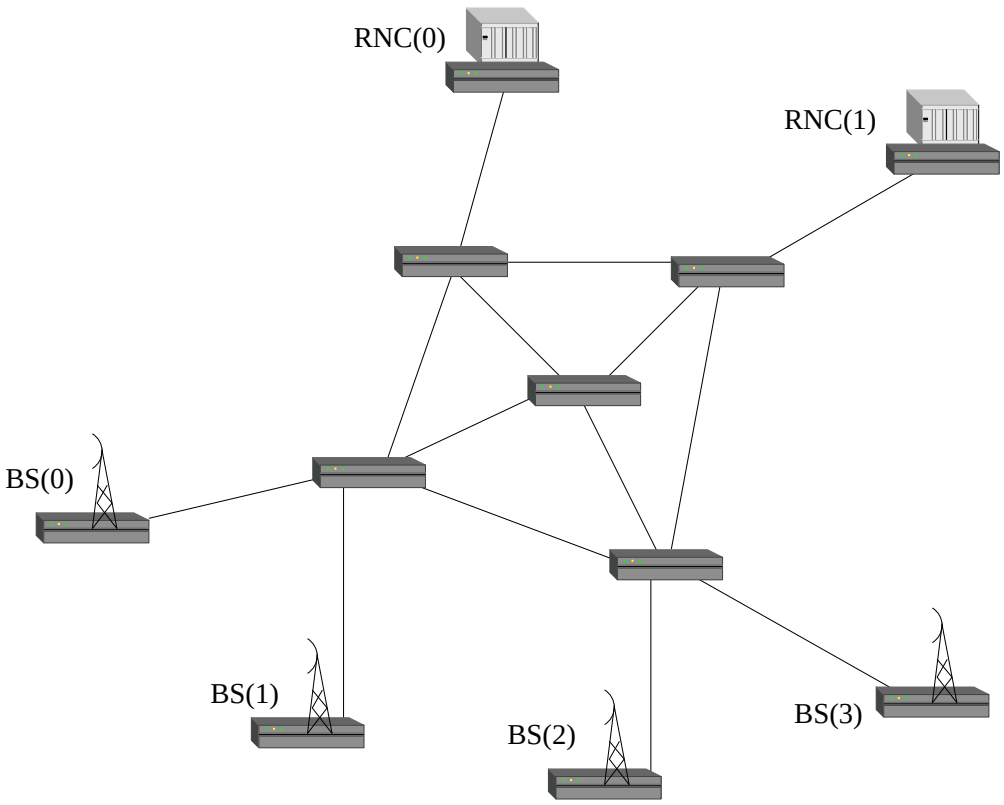


Figure 5.1: A Simple IP/MPLS-based Radio Access Network

nodes are used to construct a "IP-backbone" for traffic between BSs and RNCs. As described in our framework, in this IP/MPLS-based network, BSs and RNCs are both DiffServ ER (Edge Router) and MPLS LER (Label Edge Router); core nodes are both DiffServ CR (Core Router) and MPLS LSR (Label Switch Router)

Three types of traffic are applied to the network: user traffic (Real-Time and Non-Real-Time), control traffic, and background traffic. According to our QoS framework, *real-time user traffic* and *control plane traffic* are mapped to DiffServ EF class, *non-real-time user plane traffic* is mapped to DiffServ AF class. *Best-effort traffic*, without any bandwidth guarantees, is used as background traffic.

All three types of traffic are modeled by an On/Off source that alternates between its active and inactive period, *T_burst* and *T_idle*, respectively. *T_burst* indicates the traffic burst period; *T_idle* denotes the traffic idle period. Different path setup and tear-down algorithms are applied to different traffics according to our QoS framework.

For *user plane traffic*, both real-time traffic and non-real-time traffic, when the state changes from *idle* to *burst*, an LSP has to be set up and when the state changes from *burst* to *idle*, the path has to be torn down (release any resource being reserved, not necessary release the LSP, as there might be other traffic associated with the LSP). Hence, a user connection request is associated with its source and destination address, its bandwidth demand and the respective burst/idle periods. The QoS routing algorithm, *shortest-widest* [24] is used to find the bandwidth-constrained route. Note that in the simulation, no admission control is considered and it is assumed that the dynamical LSP setup for *user plane traffic* is never rejected, i.e., there is always enough resources to be reserved for LSP setup.

In contrast to dynamical LSP establishing and releasing for user plane traffic, the LSP for *control plane traffic* are static, or pre-established, and there is no need to set up or tear down any path when the state is changed. For *background traffic*, there is no logic connections setup or tear-down, and it is routed with common short-path-first algorithms.

In the simulation, the Layer 3 delay (including from MPLS up to Layer 3, Layer 3 packet processing, Layer 3 routing table lookup, from Layer 3 down to lower layer in the reality network stack) is modeled as a constant time delay (L3_DELAY) to differentiate the packets that have been processed in Layer 3 with the packets that are processed only in MPLS layer. The simulation parameters of interest are summarized in Table 5.2 and Table 5.3. There are 16 control connections, 16 background connections, and 32 user connections. Statistics are based on average value over all connections for each type of traffic.

5.2.3 Numerical Results

The two different DiffServ-based QoS support approaches, namely, **IP-TE** and **MPLS-TE**, are compared in terms of total network throughput, average packet delay, and per-DSCP average delay.

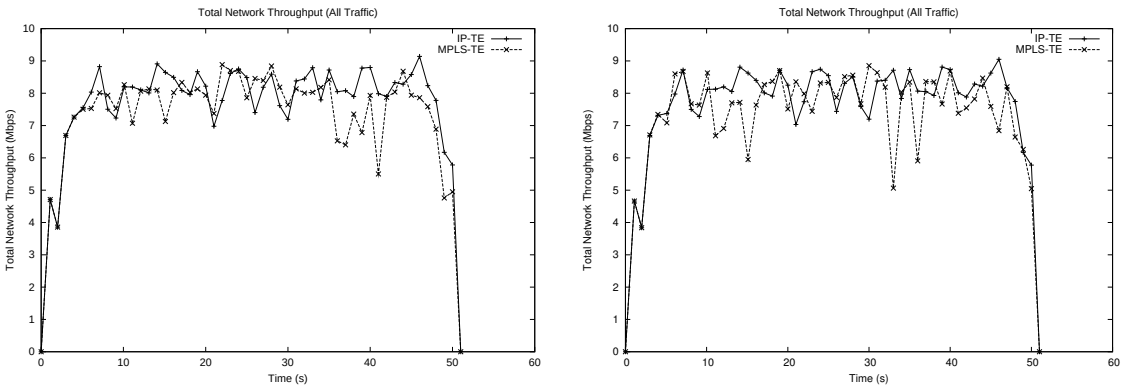
Table 5.2: Common Simulation Parameters

Simulation time	50 seconds
BS node	4
RNC nodes	2
core nodes	5
link delay	5μs
Layer 3 delay	10μs, 50μs,100μs, 1000μs
core links	8
core link bandwidth	2Mbps
User Plane Traffic	Packet Size = 100Bytes, Rate = 100Kbps
Control Plane Traffic	Packet Size = 100Bytes, Rate = 100Kbps
Background Traffic	Packet Size = uniform[0, 500Bytes], Rate = 500Kbps

Table 5.3: Traffic Simulation Parameters

Traffic	connections	Burst(s)	Idle(s)	PHB
User plane	every BS-RNC pair	uniform [2,4]	uniform [1,2]	EF
	every BS-RNC pair	uniform [2,4]	uniform [1,2]	AF
Control plane	every BS-RNC pair	uniform [2,4]	uniform [1,2]	EF
Background	every BS-RNC pair	2.0	1.0	BE

5.2.3.1 Total Network Throughput

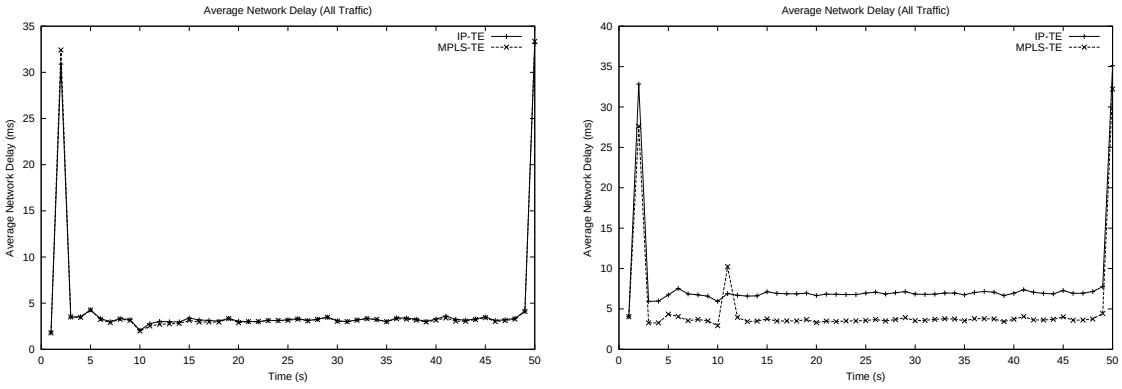


(a) L3_DELAY $= 10\mu s$ **(b)** L3_DELAY $= 1000\mu s$

Figure 5.2: Total Network Throughput for All Traffic

Figure 5.2 shows the total network throughput for all traffics, with L3_DELAY of 10μs and 1000μs. There are a few lower-throughput spikes for **MPLS-TE** around time duration 10-20s and 30-40s. The spikes become sharper for the case of L3_DELAY $= 1000\mu s$. This is due to the nature of background traffic: as background traffic is delivered by best-effort, the corresponding network throughput would be quite different and unpredictable for different L3_DELAY, which results in different spikes. Despite that, there is no significant difference between two approaches.

5.2.3.2 Average Network Delay



(a) L3_DELAY $= 10\mu s$ **(b)** L3_DELAY $= 1000\mu s$

Figure 5.3: Average Network Delay for All Traffic

Figure 5.3 shows the average packet delay for all traffics, with L3_DELAY of 10μs and 1000μs. The two big spikes near simulation time 0s and 50s are caused by the startup of simulation and the end of simulation, respectively. It is observed that there is no much difference between **IP-TE** and **MPLS-TE** for the case of L3_DELAY$= 10\mu s$ (Figure 5.3(a)); but for L3_DELAY$= 1000\mu s$ (Figure 5.3(b)), **IP-TE** leads to much larger average packet delay, while the delay performance of **MPLS-TE** is not affected. This is due to the layer 3 delay being modeled in the simulation. This can also be verified by Figure 5.4: when L3_DELAY increases from $10\mu s$ to 1000μs, there is a

"faster" increase of average network delay for **IP-TE** and "slower" increase for **MPLS-TE**.

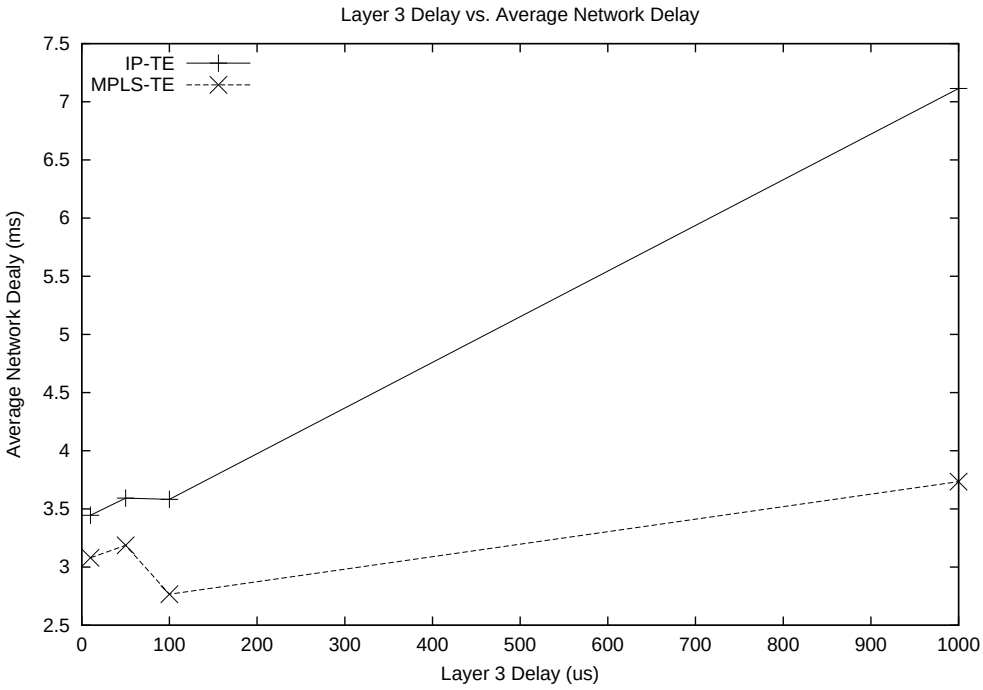
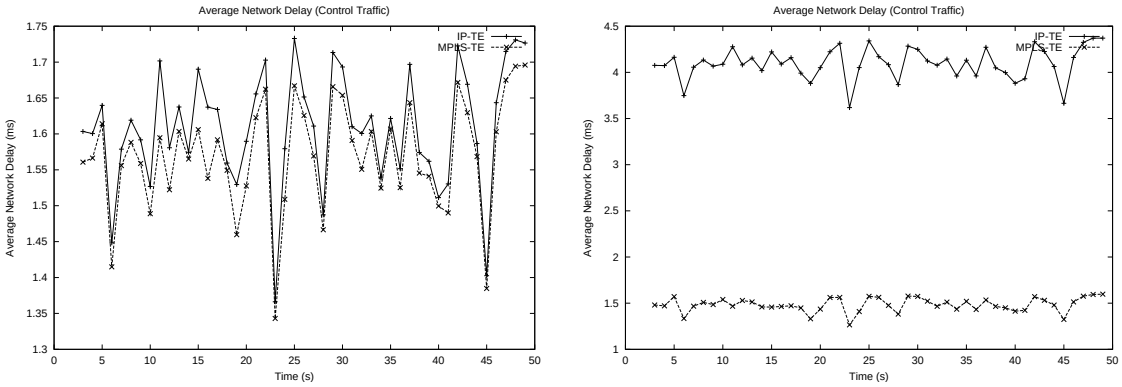


Figure 5.4: Layer 3 Delay vs. Average Network Delay for All Traffic

For **MPLS-TE** with L3_DELAY$= 1000\mu s$ (Figure 5.3(b)), it is observed that there is a small spike near simulation time 10s. This is due to the setup delay of LSP for user plane traffic as all traffic is started at simulation time 10s in the simulation. When L3_DELAY is not large comparing with traffic load (Table 5.2), such setup delay is not significant and hence there is no spike with L3_DELAY$= 10\mu s$ (Figure 5.3(a)). This can be verified with a more in-depth examination of average packet delay for the three different traffic used in the simulation, i.e., *control plane traffic* (Figure 5.5), *user plane traffic* (Figure 5.6), and *background traffic* (Figure 5.7). There is no such spike for *control plane traffic* and *background traffic*, as LSPs for control plane traffic are pre-established and no LSP is setup for background traffic. For *user plane traffic*, there is a sharp spike when L3_DELAY is large (Figure 5.6(b)) as LSP for user plane traffic are dynamically established.

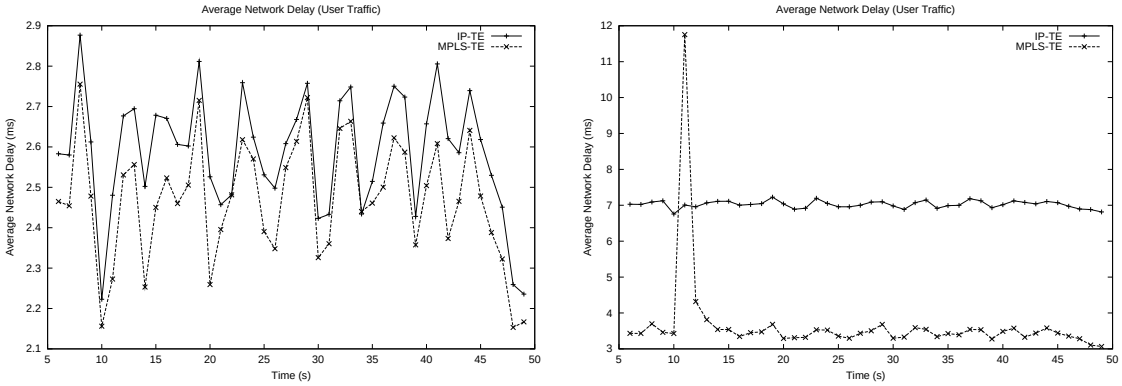
It is observed that, in Figure 5.4, **MPLS-TE** curve shows a decreasing trend just

before 100μs. This is due to the way the simulation is carried out: traffic is not injected into the network at the same time and hence different LSP setup delay is incurred, which introduces different simulation startup delay for **MPLS-TE**. That is why, shown in Figure 5.3, a bigger spike around simulation time 3s is observed with L3_DELAY= 10μs than L3_DELAY= 1000μs for **MPLS-TE**, but similar spike for **IP-TE**.



(a) L3_DELAY = 10μs **(b)** L3_DELAY = 1000μs

Figure 5.5: Average Network Delay for Control Plane Traffic



(a) L3_DELAY = 10μs **(b)** L3_DELAY = 1000μs

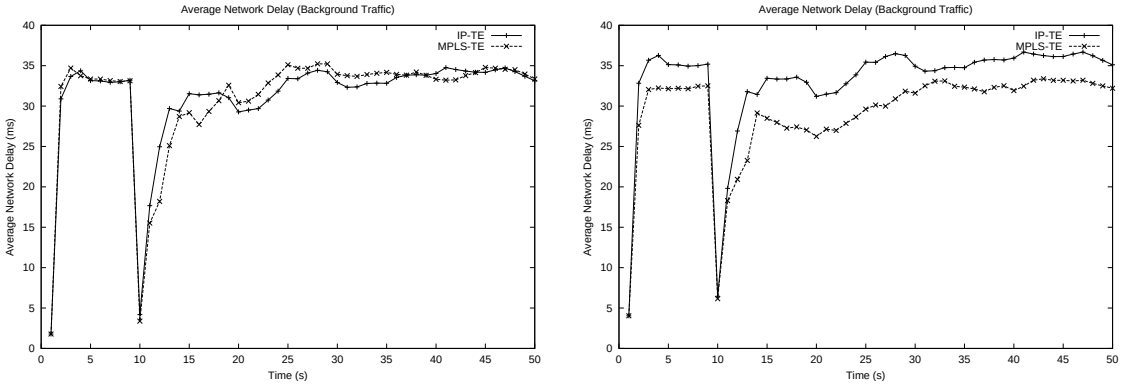
Figure 5.6: Average Network Delay for User Plane Traffic

Besides the startup LSP-setup-delay spike, it is observed that there is no other such spike during the simulation. This suggests that those dynamic LSPs are never released

after being established when traffic start, which might be a unfair comparison for **IP-TE**. But one reason for that is due to the LSP setup mechanism adopted in the QoS framework (Chapter 3, Section 3.2): instead of establishing LSP for every flow of traffic, LSPs are established between RNC and BSs for all traffic and resources (e.g. bandwidth) are reserved aggregated, or in the basis of LSP. LSPs are released only when all resources (e.g. bandwidth) being reserved have been released. In the simulation, due to the traffic pattern simulated, there is no occasion when all traffic on one LSP has been in *T_idle* state. Consequently there is no LSP being released and no more LSP-setup-delay spikes. Although some delay will be incurred due to signaling along the LSP in order to adjust the bandwidth reservation whenever a flow joins or leaves, it does not affect the overall delay performance as only short link delay (Table 5.2) is involved for resource signaling along the LSP.

5.2.3.3 Per-DSCP Average Delay

Table 5.4 and 5.5 show the per-DSCP average packet delay for **IP-TE** and **MPLS-TE**, respectively. It is observed that for EF (DSCP 46) and AF (DSCP 10), both approaches have achieved good performance for delay, while **MPLS-TE** results in better delay performance for AF and BE. Moreover, shown in Figure 5.8 (the per-DSCP average



(a) L3_DELAY $= 10\mu s$ **(b)** L3_DELAY $= 1000\mu s$

Figure 5.7: Average Network Delay for Background Traffic

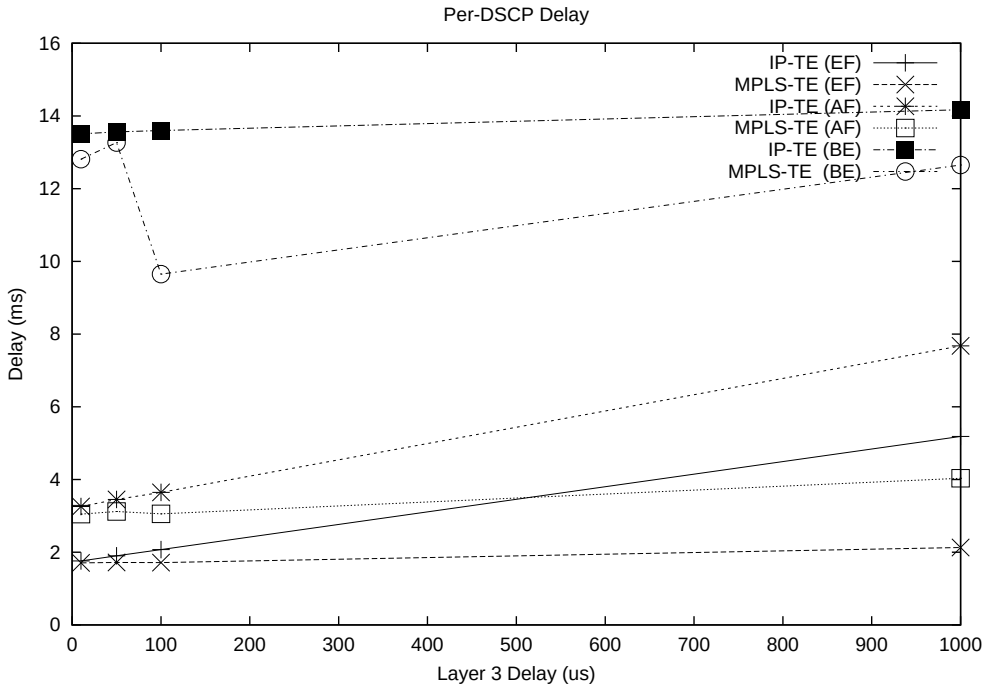


Figure 5.8: Per-DSCP Average Network Delay

network delay with increasing L3_DELAY), **MPLS-TE** leads to better delay performance for all DSCPs. In addition, it is observed that maximum delay (max-delay) for DSCP=0 in **MPLS-TE** is much larger than that in **IP-TE**. This is due to the treatment for traffic with DSCP=0 and LSP setup delay in **MPLS-TE**. As there is no resource reservation for background traffic (DSCP=0), they are delivered by best-effort. When there are long LSP setup delay in **MPLS-TE**, long delay may be caused for background traffic; while in **IP-TE**, there is no LSP setup at all and background traffic would not be affected, though still by best-effort.

Table 5.4: IP-TE: Per-DSCP average delay

DSCP	min-delay(ms)	max-delay(ms)	mean-delay(ms)
46	0.595	21.7	1.761
10	1.425	228	3.258
0	0.021	15.01	13.51

Table 5.5: MPLS-TE: Per-DSCP average delay

DSCP	min-delay(ms)	max-delay(ms)	mean-delay(ms)
46	0.051	21.49	1.706
10	1.395	184.3	3.049
0	0.021	166.3	12.81

5.3 Simulation on IP-RAN Micro-Mobility

A simplified Hierarchical Mobile IPv6 ([37]) model was constructed in ns-2 to study three different tunnel-based micromobility schemes, namely **IP-Tunnel**, **MPLS-Tunnel**, and **MPLS-Tunnel with LMA** (Chapter 4, Section 4.3). Performance are compared with respect to some important handover performance metrics, such as handover latency, packet loss ratio, and average packet delay.

- **IP-Tunnel**: IP Tunnel, or IP encapsulation [33], is used for packet forwarding. Upon receiving BU message from a mobile host, RNC encapsulates the packets to the mobile host towards the BS where the mobile host is currently attached, with additional IP header (20 Bytes for IPv4 [26] and 40 Bytes for IPv6 [54]). Upon receiving BU message from a mobile host after handover, RNC encapsulates the packets with the new CoA towards the new BS. This scheme may cause the problem of added overhead, especially when the size of the data packets is comparable with the size of IP header, but it is very simple.

- **MPLS-Tunnel**: MPLS Tunnel, or LSP tunnel [27], is used for packet forwarding. LSP tunnels are established between RNC and BSs. Upon receiving BU message from a mobile host, RNC labels the packets to the mobile host and use the LSP to the corresponding BS to forward the packet, with additional MPLS header (4 Bytes [27]). Upon receiving BU message from a mobile host after handover, RNC redirects the packets to the LSP to the new BS. This scheme may add less overhead, compared to **IP-Tunnel**, but additonal complexity is needed for normal

MPLS functionality for all network elements.

- **MPLS-Tunnel with LMA**: MPLS Tunnel, or LSP tunnel [27], is also used for packet forwarding. LSP tunnels are established between RNC, LMA, and BSs. Upon receiving BU message from a mobile host, RNC/LMA labels the packets to the mobile host and use an LSP to the corresponding BS to forward the packet, also with additional MPLS header (4 Bytes [27]). Upon receiving BU message from a mobile host after handover, RNC/LMA redirects the packets to the LSP to the new BS. As LMA can redirect the LSP before BU reaches RNC, only partial LSP may be needed to be redirected, compared to **MPLS-Tunnel**. This scheme may reduce handover latency, compared to both **MPLS-Tunnel** and **IP-Tunnel**, but additional complexity is needed for normal MPLS functionality for all network elements and LMA functionality for some network elements.

The network topology shown in Figure 5.9 is used in the simulation. Such a network topology is simple but yet good enough to investigate various handover performance issues, such as handover latency and packet loss ratio, etc. There are 4 Base Stations, and 1 RNC/Gateway, and 5 RAN router. The *ran*(3) and *ran*(1) can be identified as LMA, following the methods described in Section 4.1.2.2.

Due to the lack of simulation model for UTRAN/UMTS, as well as WCDMA air interface and related MAC protocol, ns-2 IEEE 802.11 wireless LAN model is used for the mobile host to connect with the base station. As the main objective of the simulation study is to investigate the handover performance issues related with location management, i.e., location registration, it is reasonable to use such a model although it is not realistic. Consequently, some parameters related to simulation setup might not be very meaningful, e.g., the overlap distance of base station, comparing to usual wireless LAN simulation. For simplicity, only two-direction movements are considered for the mobile hosts in the simualtions, i.e., the mobile host moves either left or right.

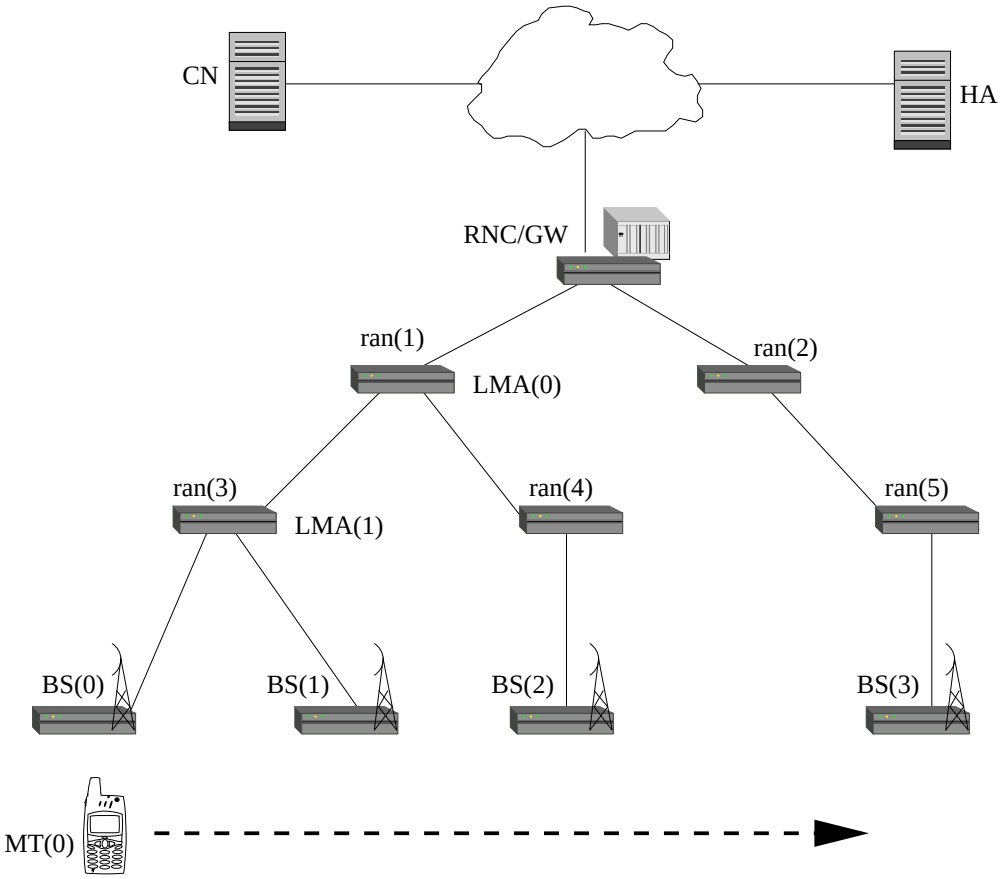


Figure 5.9: Micro-Mobility Simulation Network Topology

A simple handover mechanism is used in the simulation. When the mobile host is moving towards the cell covered by other base stations, upon receiving the first beacon signal from the new base station, the mobile host assumes that a handover has occurred and notified the base station. In addition, once the mobile host initiates a handover to a new base station, it is not able to receive packet (except for broadcast beacons) from the old base station.

5.3.1 CBR Traffic

Constant Bit Rate (CBR) traffic is used firstly to verify how the three schemes work. The common parameters for simulation are shown in Table 5.6. Following the network topology shown in Figure 5.9, a simple mobility model is used for the mobile host: staring from the cell of Base Station BS(0), and moving towards to Base Station BS(3),

Table 5.6: Common Simulation Parameters: CBR Traffic

Simulation time	600 seconds
Layer 3 delay	0
BS range	250 m
Overlap of BS	0 m
Number of MH	1
Speed of MH	uniform [5, 15] m/s
RAN link bandwidth	2 Mbps
RAN link delay	5 μs - 10ms
Packet Size	100 Bytes
Traffic Interval	10 ms

passing Base Station BS(1) and BS(2), and then backward to BS(0), and then toward to BS(3), and so on.

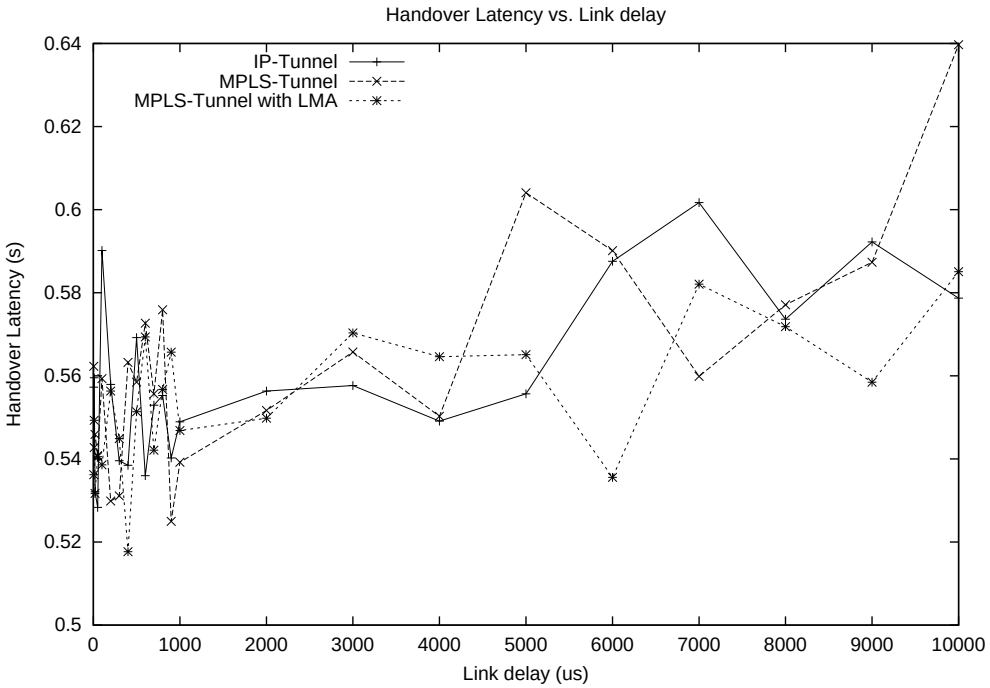


Figure 5.10: Handover Latency vs. Link Delay

For *handover latency* (Figure 5.10), it is observed that all three schemes fluctuate when link delay is smaller than 5ms; when link delay is larger than 5ms, the difference between three schemes get more obvious. **MPLS-Tunnel with LMA** does not

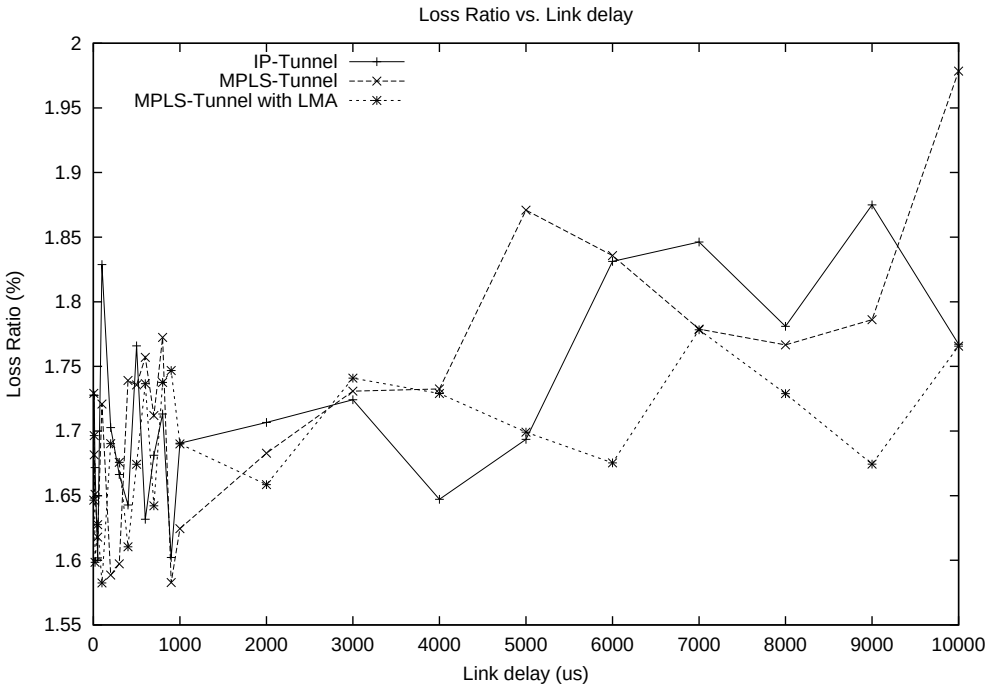


Figure 5.11: Packet Loss Ratio vs. Link Delay

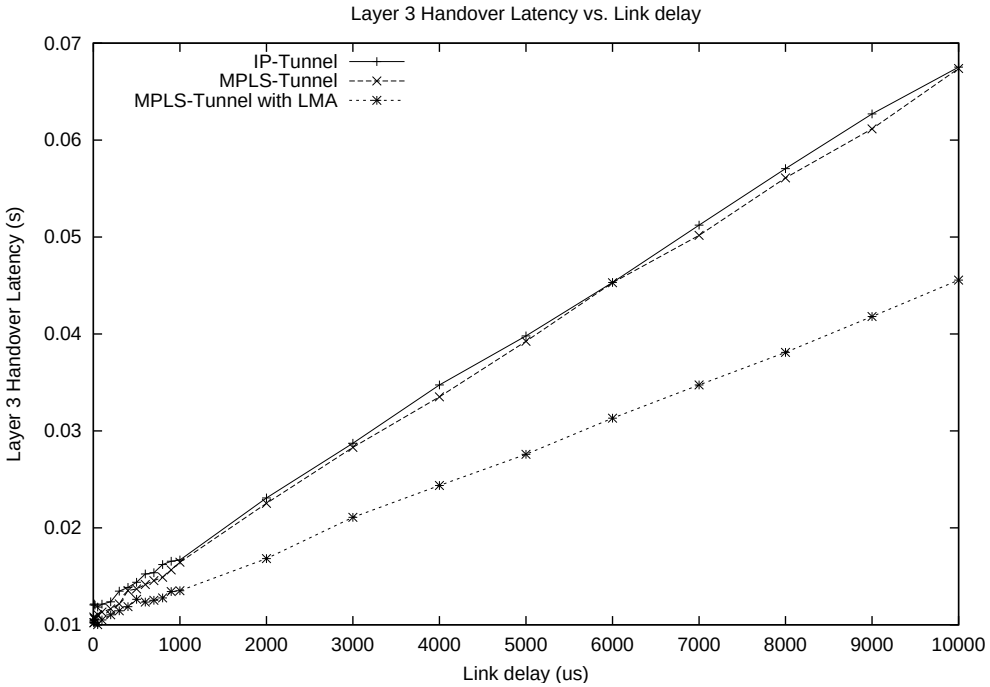


Figure 5.12: Layer 3 Handover Latency vs. Link Delay

perform better that the other two schemes. Similar observation can be obtained for

packet loss ratio (Figure 5.11), except that **MPLS-Tunnel with LMA** do result in better performance when link delay is larger than 5ms. These seem to be different from what is expected. Given the network topology (Figure 5.9), when MH moves from the cell of BS(0) to BS(1), *ran*(3)/LMA(1) is able to identify the handover by snooping the BU message sent by the mobile host (Section 4.3.2). LMA(1) then can redirect immediately the packets for the mobile host from the LSP to BS(0) to the LSP to BS(1), without waiting the BU to reach the RNC/Gateway. Thus for **MPLS-Tunnel with LMA**, it is expected that *handover latency* and *packet loss ratio* should be reduced compared to the other two schemes.

Handover latency is defined for a receiving mobile host as the time that elapses between the last packet received via the old route and the arrival of the first packet along the new route after a handover [55]. This time delay can be separated into two parts in the simulation, *Layer 2 Handover Latency* and *Layer 3 Handover Latency*. *Layer 2 Handover Latency* includes the new IP address prefix discovery on the new IP subnet, the new CoA address establishment, whereas *Layer 3 Handover Latency* refers to the time needed to notify RNC/Gateway. Comparing to **IP-Tunnel** and **MPLS-Tunnel**, **MPLS-Tunnel with LMA** does not reduce *Layer 2 Handover Latency*, but only *Layer 3 Handover Latency*.

With more detail examination of the simulation setting, it is found that the unexpected results with *handover latency* and *packet loss ratio* are due to the handover mechanism used in the simulation, which introduces different setting, or unfair comparison, for simulation with different schemes. *Layer 2 handover latency*, is different for the three schemes, thus the results are not as expected. This can be verified to see how *Layer 3 Handover Latency* varies with the increase of link delay, shown in Figure 5.12. It is observed that both **IP-Tunnel** and **MPLS-Tunnel** result in similar *Layer 3 handover latency*, and **MPLS-Tunnel with LMA** does reduce *Layer 3 handover latency* with the increase of link delay. To avoid this problem caused by simulation setting and to have a fair comparison among the three schemes, simulation setting is modified

accordingly for the subsequent simulations.

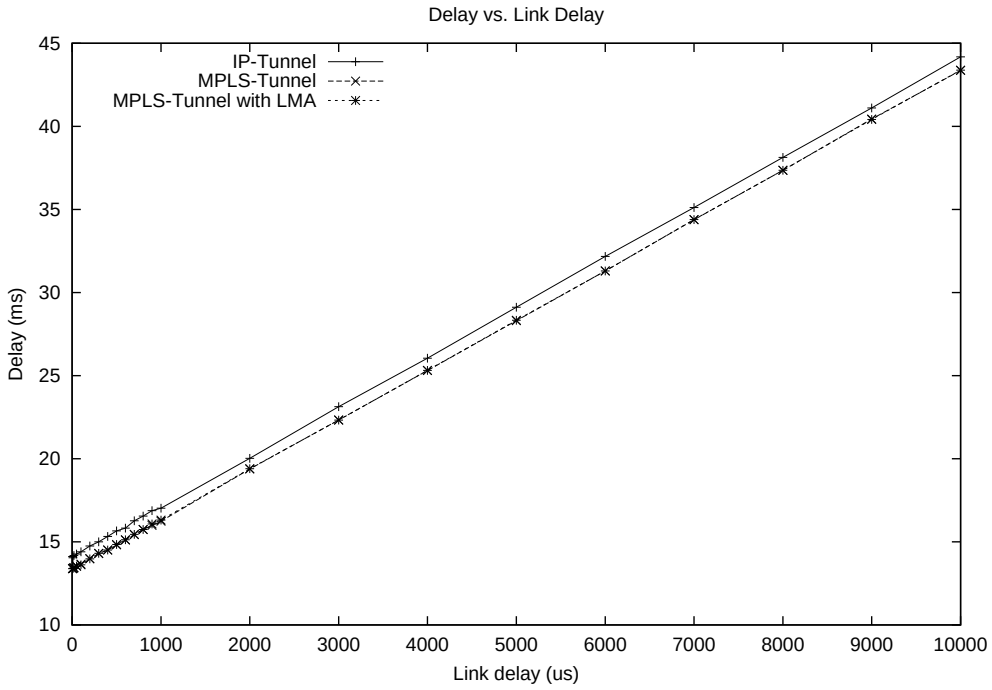


Figure 5.13: Average Packet Delay vs. Link Delay

For *average packet delay* (Figure 5.13), **MPLS-Tunnel** and **MPLS-Tunnel with LMA** result in the same performance, while **IP-Tunnel** results in larger delay due to IP encapsulation overhead.

5.3.2 ON/OFF Traffic

ON/OFF traffic are used to model the RAN traffic: data packets are sent only during ON period, and no data packets during OFF period. Two scenarios, namely variation of mobility patterns and variation of traffic rate, are carried out. For simplicity, three kinds of mobility are used in simulation to model the movement of mobile hosts in radio access networks: no movement ($speed = 0$), low mobility ($0 < speed < 10m/s$), fast mobility ($20 < speed < 30m/s$).

5.3.2.1 Variation of Mobility Patterns

In this scenario, three mobility patterns are considered for one mobile host. Given the network topology (Figure 5.9), there are three crossover routers, corresponding to different crossover distances: RNC, three-hop away from all BSs; LMA(0), two-hop away from BS(0), BS(1) and BS(2); LMA(1), one-hop away from BS(0) and BS(1).

- mobility pattern 1: the mobile host moves forwards and backwards between BS(0) and BS(1) with different speed at different time instances, i.e., from BS(0) to BS(1) and then BS(1) to BS(0), and so on. In this case, there is one crossover router involved, LMA(1).

- mobility pattern 2: the mobile host is moving forwards and backwards among BS(0) and BS(2) with different speed at different time instances, i.e., from BS(0) to BS(1), to BS(2), and then from BS(2) to BS(1), to BS(0), and so on. In this case, there is two crossover routers involved, LMA(1) and LMA(0).

- mobility pattern 3: the mobile host is moving forwards and backwards among BS(0) and BS(3) with different speed at different time instances, i.e., from BS(0) to BS(1), to BS(2), to BS(3), and then from BS(3) to BS(2), to BS(1), to BS(0), and so on. In this case, there is three crossover routers involved, LMA(1), LMA(0) and RNC.

The common parameters for simulation are shown in Table 5.7.

Figure 5.14, 5.15 and 5.16 show the results for *handover latency*, *packet loss ratio* and *average packet delay*, respectively. Among three mobility patterns, it is observed that the three schemes result in different *handover latency* and *packet loss ratio*, but the same *average packet delay*. Moreover, larger *handover latency* leads to bigger *packet loss ratio*.

Table 5.7: Common Simulation Parameters: Variation of Mobility Patterns

Simulation time	3600 seconds
Layer 3 delay	0
BS range	500 m
Overlap of BS	0 m
Number of MH	1
Speed of MH	three patterns, 0 - 30 m/s
RAN link bandwidth	2 Mbps
RAN link delay	6ms
Packet Size	100 Bytes
Call ON time	1.004 s
Call OFF time	1.587 s
Traffic Rate	100 kbps

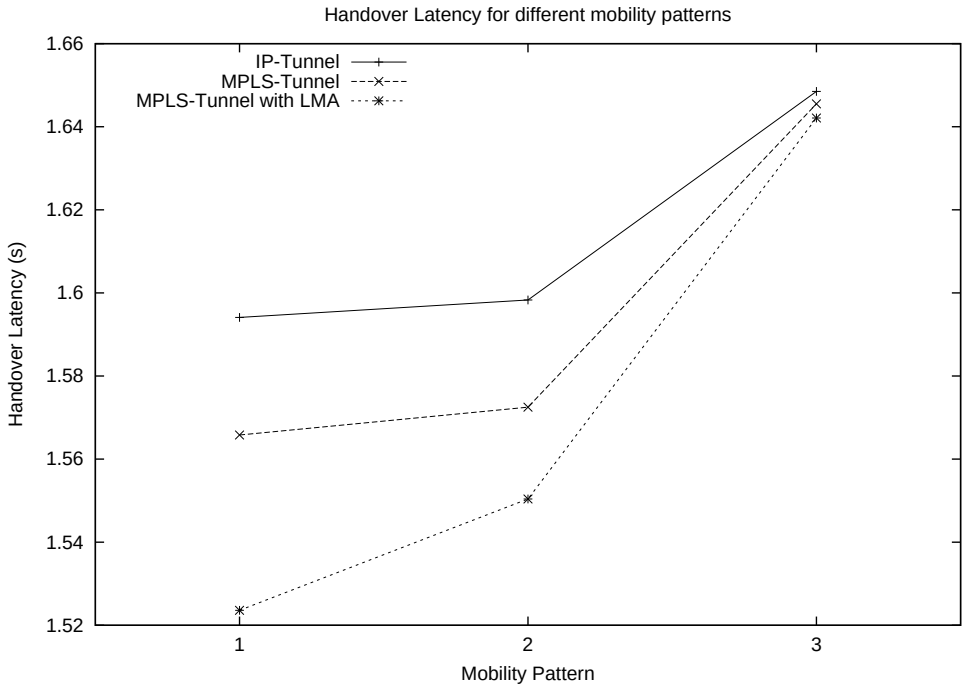


Figure 5.14: Handover Latency for different mobility patterns

Comparing different schemes for different mobility patterns, **MPLS-Tunnel with LMA** results in the smallest *handover latency* (Figure 5.14) and *packet loss ratio* (Figure 5.15) for all three mobility patterns. As expected, this is due to partial LSP-redirection achieved by **MPLS-Tunnel with LMA**. For the remaining two schemes,

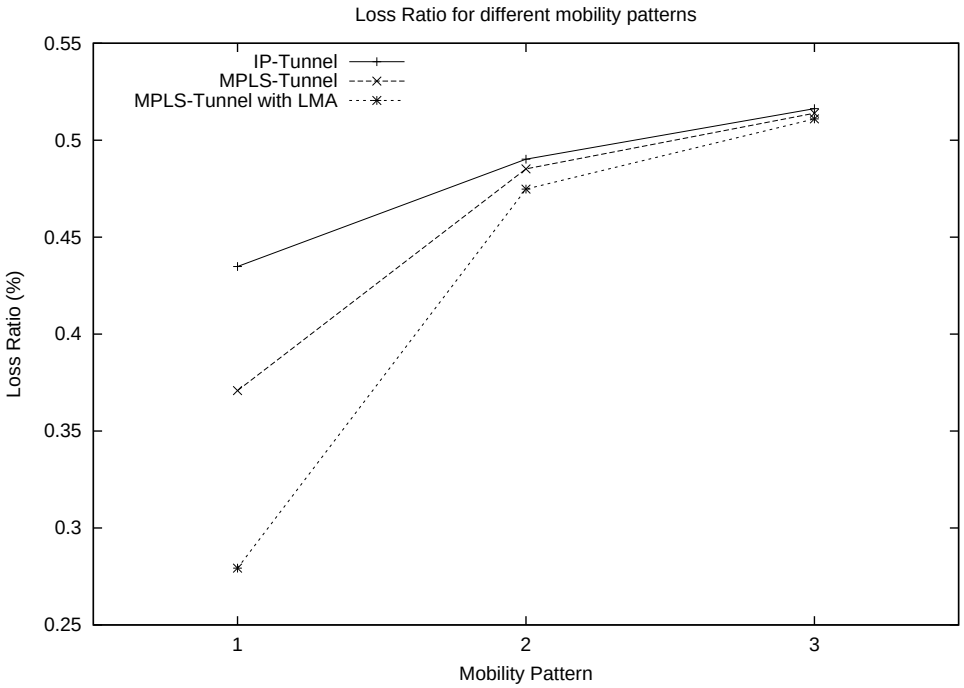


Figure 5.15: Packet Loss Ratio for different mobility patterns

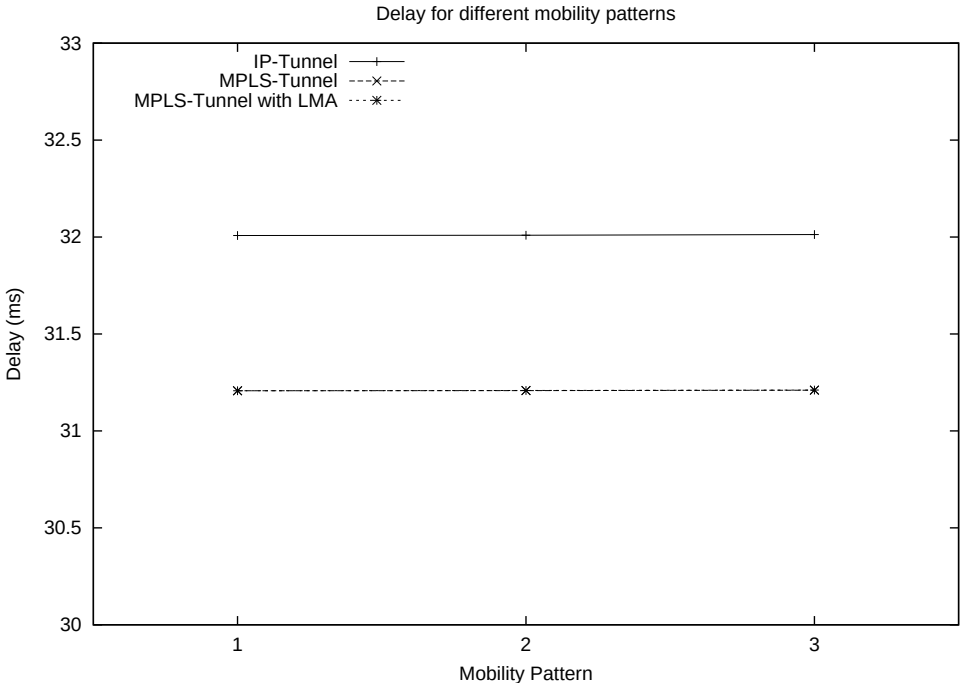


Figure 5.16: Average Packet Delay for different mobility patterns

due to the encapsulation overhead, **IP-Tunnel** results in larger *handover latency* and

packet loss ratio than **MPLS-Tunnel**. This is also observed from Figure 5.16, **IP-Tunnel** results in larger constant *average packet delay* than the other two MPLS-based schemed.

Moreover, the best handover performance, in terms of *handover latency* and *packet loss ratio*, is obtained for mobility pattern 1. mobility pattern 2 has worse performance comparing to mobility pattern 1, but better performance comparing to mobility pattern 3. As **MPLS-Tunnel with LMA** would achieve better handover performance if the crossover router is nearer the BS, which is the case for mobility pattern 1.

5.3.2.2 ON/OFF Traffic: Variation of Traffic Rate

In this scenario, traffic rate is varied from 10kbps to 140kbs and link delay is fixed to 5ms, to model increasing network traffic load in the RAN. The common parameters for simulation are shown in Table 5.8. The results for *handover latency*, *packet loss ratio*, and *average packet delay* are shown in Figure 5.17, 5.18 and 5.19 , respectively.

Table 5.8: Common Simulation Parameters: Variation of Traffic Rate

Simulation time	600 seconds
Layer 3 delay	0
BS range	500 m
Overlap of BS	0 m
Number of MH	20
Speed of MH	three patterns, 0 - 30 m/s
RAN link bandwidth	2 Mbps
RAN link delay	50 μs - 10ms
Packet Size	100 Bytes
Call ON time	1.004 s
Call OFF time	1.587 s
Traffic Rate	10 kbps - 140 kbps

Observed from Figure 5.17, when traffic rate is greater than 80kbps, **MPLS-Tunnel with LMA** results in smaller *handover latency* than **IP-Tunnel** and **MPLS-Tunnel**. As

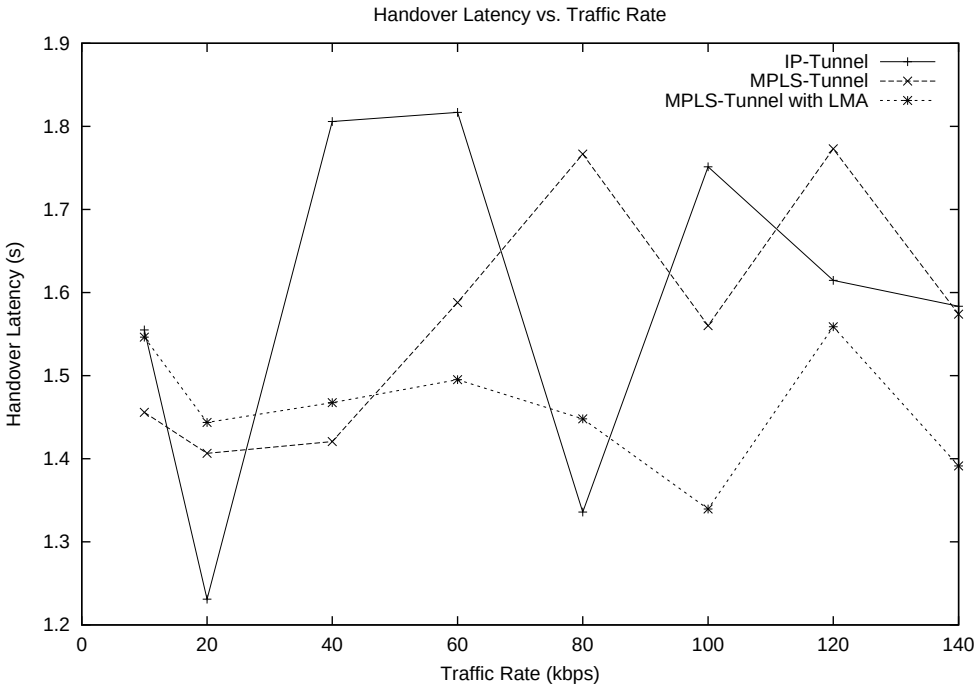


Figure 5.17: Handover Latency vs. Traffic Rate (Link Delay: 5ms)

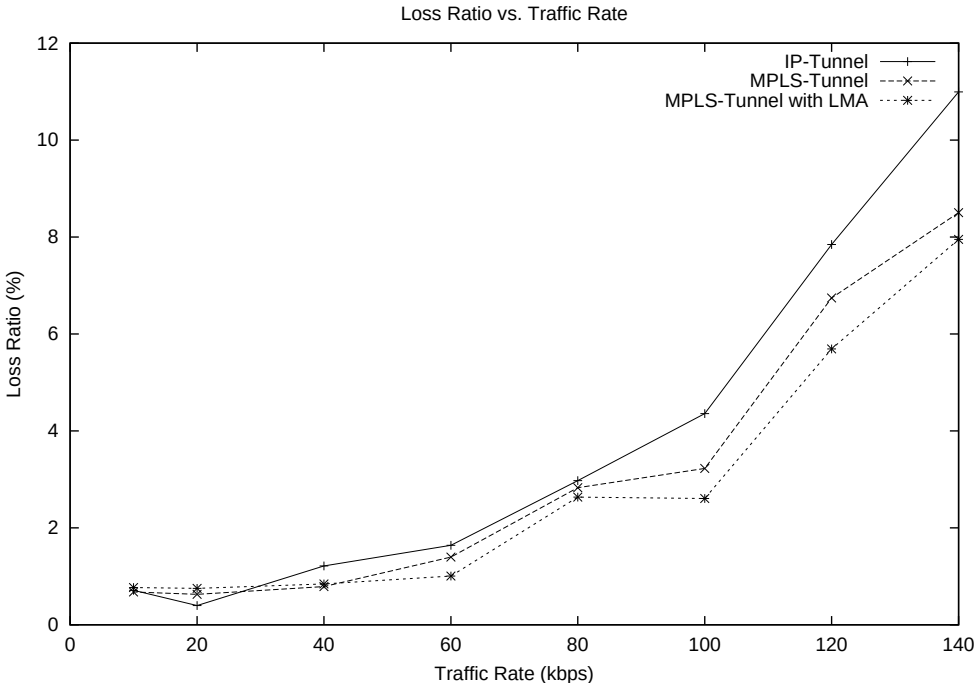


Figure 5.18: Packet Loss Ratio vs. Traffic Rate (Link Delay: 5ms)

expected, as traffic rate of 80kbps is corresponding to packet arriving interval of 10ms, when traffic rate increases, packet arrives in smaller interval during ON time and hence

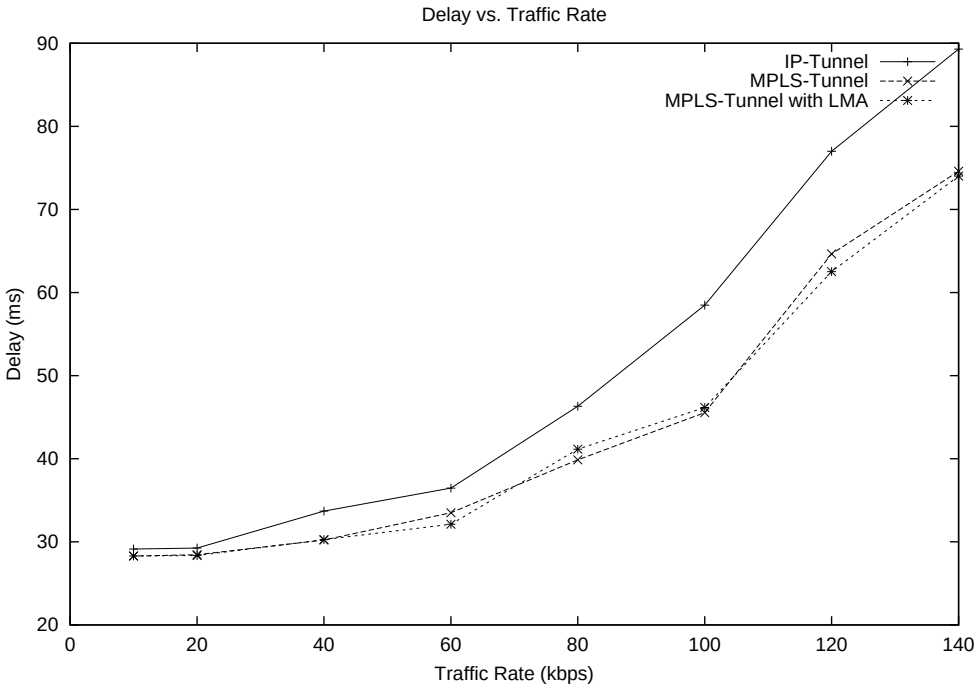


Figure 5.19: Average Packet Delay vs. Traffic Rate (Link Delay: 5ms)

MPLS-Tunnel with LMA reduces the time the packet be forwarded. consequently a lower packet loss ratio is achieved (Figure 5.18).

For *packet loss ratio* (Figure 5.18), all three schemes lead to comparable loss performance when traffic rate is small. When traffic rate, or network load, increases, the two MPLS-based schemes achieve better performance, and **MPLS-Tunnel with LMA** leads to the best performance.

For *average packet delay* (Figure 5.19), **IP-Tunnel** results in larger packet loss ratio than the other two MPLS-based schemes. This is due to the overhead of IP encapsulation, as the average packet delay for **IP-Tunnel** is the largest.

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

A framework for QoS support in IP/MPLS-based radio access network is presented. Two different approaches for Traffic Engineering in radio access networks under the same DiffServ QoS model, namely IP-TE and MPLS-TE, are compared through simulation study. The simulation results, including total network throughput, average network delay, and Per-DSCP average delay, show that the proposed approach does satisfy the diverse QoS requirements while efficiently utilizing network resources in the RANs. In addition, MPLS-TE results in better delay performance, which is exactly the most important transport requirement for IP-RAN.

A hierarchical-based micro-mobility scheme integrated with QoS for IP/MPLS-based 3G RANs is proposed. By introducing one more level of hierarchy in hierarchical-based radio access network and locating the Local Mobility Agent in the network, two-stage LSPs can be established to take advantage of common LSP path during handover: only partial LSP re-direction is needed for handover and hence the handover latency is reduced. The interoperation of the micro-mobility scheme with Hierarchical Mobile IPv6 is described and simulation results show that this approach improve handover performance, in terms of handover latency and packet loss.

6.2 Future Work

While the work in this thesis only concerns some issues in IP-based radio access networks, there are a few more issues that could be addressed:

1. The micro-mobility simulation in Chapter 5 Section 5.3 only consider one class

of traffic. Some work could be done to investigate the performance for multiple classes of traffic, i.e., integration of QoS and Micro-Mobility.

2. For LMA operation in the micromobility scheme in Chapter 4, a LMA Field is proposed to be added in MIPv6 BU message for LMAs to update its address. It might be useful to add the field in MIPv6 BACK message so that the mobile host could know its attaching LMA for possible handover optimization.

3. Integration of micro-mobility with macro-mobility management would be a challenge yet meaningful work. Various micro-mobility schemes have been proposed, but they are usually studied only in a micro-scope point of view. It is worthwhile to investigate the different tradeoff and impact from protocol design in a macro-scope point of view.

REFERENCES

[1] MWIF, "OpenRAN Architecture in 3rd Generation Mobile Systems," Mobile Wireless Internet Forum (MWIF)," Technical Report MTR-007, Release v1.0.0, Sept. 2001. [Online]. Available: http://www.mwif.org

[2] N. Gerlich and H. Becker, "A Functional Architecture for 3G IP based Radio Access Networks," in *2001 International Conference on Third Generation Wireless and Beyond (3Gwireless'01)*, San Francisco, USA.

[3] D. Awduche et al., "Overview and Principles of Internet Traffic Engineering," RFC 3272, May 2002.

[4] Y. Wang and Z. Wang, "Explicit Routing Algorithms for Internet Traffic Engineering," in *Eight International Conference on Computer Communications and Networks (ICCCN'99)*, Oct. 1999, pp. 582–588.

[5] Andrew T. Campbell et al., "Comparison of IP Micromobility Protocols," *IEEE Wireless Communications Magazine*, vol. 9, no. 2, pp. 2–12, Feb. 2002.

[6] Feng Li, Hoang M. Nguyen and Winston K.G. Seah, "QoS Support in IP/MPLS-based Radio Access Networks," in *IEEE 57th Vehicular Technology Conference (VTC Spring 2003)*, Jeju, Korea, Apr. 2003.

[7] Hoang M. Nguyen, Feng Li and Queying Xie, "Integration of Micro-Mobility with QoS in IP/MPLS-Based Radio Access Networks," in *IEEE 57th Vehicular Technology Conference (VTC Spring 2003)*, Jeju, Korea, Apr. 2003.

[8] S. Dixit, Y. Guo, and Z. Antoniou, "Resource Management and Quality of Service in Third-Generation Wireless Networks," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 125–133, Feb. 2001.

[9] 3GPP TS 23.002, "Network Architecture," rel.5, v. 5.6.0, Mar. 2002. [Online]. Available: http://www.3gpp.org

[10] F. M. Chiussi, D. A. Khotimsky, and S. Krishnan, "Mobility Management in Third-Generation All-IP Networks," *IEEE Communications Magazine*, vol. 40, no. 9, pp. 2–13, Sept. 2002.

[11] 3GPP TS 25.401, "UTRAN Overall Description," rel.5, v. 5.2.0, Mar. 2002. [Online]. Available: http://www.3gpp.org

[12] J. Kempf and P. Yegani, "OpenRAN: A New Architecture for Mobile Wireless Internet Radio Access Networks," *IEEE Communications Magazine*, vol. 40, no. 5, pp. 118–123, May 2002.

[13] J. H. Saltzer, D.P. Reed and D.D. Clark, “End-to-end Arguments in System Design,” *ACM Transactions in Computer Systems*, vol. 2, no. 4, pp. 277–288, Nov. 1984.

[14] Y. Bernet, “The Complementary Roles of RSVP and Differentiated Services in the Full-Service QoS Network,” *IEEE Communications Magazine*, vol. 38, no. 2, pp. 154–162, Feb. 2000.

[15] R. Braden, D. Clark, and S. Shenker, “Integrated Services in the Internet Architecture: an Overview,” RFC 1633, June 1994.

[16] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, “Resource Reservation Protocol (RSVP) Version 1 Functional Specification,” RFC 2205, Sept. 1997.

[17] S. Blake et al., “An Architecture for Differentiated Services,” RFC 2475, Dec. 1998.

[18] S. Shenker, C. Partridge, and R. Guerin, “Specification of Guaranteed Quality of Service,” RFC 2212, Sept. 1997.

[19] J. Wroclawski, “Specification of the Controlled Load Quality of Service,” RFC 2211, Sept. 1997.

[20] K. Nichols, S. Blake, F. Baker, and D. Black, “Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers,” RFC 2474, Dec. 1998.

[21] B. Davie et al., “An Expedited Forwarding PHB (Per-Hop Behavior),” RFC 3246, Mar. 2002.

[22] J. Heinanen, F. Baker, W. Weiss, and J. Wrocklawski, “Assured Forwarding PHB Group,” RFC 2597, June 1999.

[23] E. Crawley et al., “A Framework for QoS-based Routing in the Internet,” RFC 2386, Aug. 1998.

[24] Z. Wang and J. Crowcrof, “Quality-of-Service Routing for Supporting Multimedia Applications,” *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 7, pp. 1228–1234, Sept. 1996.

[25] J. Moy, “OSPF Version 2,” RFC 2328, Apr. 1998.

[26] J. Postel, “Internet Protocol,” RFC 791, Sept. 1981.

[27] E. Rosen, A. Viswanathan, and R. Callon, “Multiprotocol Label Switching Architecture,” RFC 3031, Jan. 2001.

[28] T. Li and Y. Rekhter, “Provider Architecture for Differentiated Services and Traffic Engineering (PASTE),” RFC 2430, Oct. 1998.

[29] D. Awduche et al., “Requirements for Traffic Engineering over MPLS,” RFC 2702, Sept. 1999.

[30] Jukka Manner et al., “Evaluation of Mobility and Quality of Service Interaction,” *Computer Networks*, vol. 38, no. 2, pp. 137–163, Feb. 2002.

[31] C. Perkins (Editor), “IP Mobility Support for IPv4,” RFC 3344, Aug. 2002.

[32] D. Johnson, C. Perkins, and J. Arkko, “Mobility Support for IPv6,” Internet-Draft, work in progress, Feb. 2003.

[33] C. Perkins (Editor), “IP encapsulation within IP,” RFC 2003, Oct. 1996.

[34] Andrew T. Campbell et al., “Design, Implementation and Evaluation of Cellular IP,” *IEEE Personal Communications Magazine*, vol. 7, no. 8, pp. 42–49, Aug. 2000.

[35] Rammchandran Ramjee et al., “HAWAII: A Domain-Based Approach for Supporting Mobility for Supporting Mobility in Wide-Area Wireless Networks,” *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 396–410, June 2002.

[36] E. Gustafsson, A. Jonsson, and C. Perkins, “Mobile IP Regional Registration,” Internet draft, work in progress, Mar. 2002.

[37] H. Soliman et al., “Hierarchical Mobile IPv6 Mobility Management (HMIPv6),” Internet draft, work in progress, June 2003.

[38] K. Venken, D. D. Vleeschauwer, and J. D. Vriendt, “Designing A DiffServ-Capable IP-Backbone For The UTRAN,” in *Second International Conference on 3G Mobile Communication Technologies*, London, UK, Mar. 2001, pp. 47–52.

[39] Y. Guo, Z. Antoniou, and S. Dixit, “IP Transport in 3G Radio Access Networks: an MPLS-based Approach,” in *IEEE Wireless Communications and Networking Conference (WCNC2002)*, vol. 1, Orlando, FL, USA, Mar. 2002, pp. 11–17.

[40] F. M. Chiussi, D. Khotimsky, and S. Krishnan, “A Network Architecture for MPLS-based Micro-Mobility,” in *IEEE Wireless Communications and Networking Conference (WCNC2002)*, vol. 1, Orlando, FL, USA, Mar. 2002, pp. 549–555.

[41] Zhong Ren et al., “Integration of Mobile IP and Multi-Protocol Label Switching,” in *IEEE International Conference on Communications (ICC’01)*, vol. 7, Helsinki, Finland, June 2001, pp. 2123–2127.

[42] G. Eneroth, G. Fodor, G. Leijonhufvud, A. R*ácz*, and I. Szab*ó*, “Applying ATM/AAL2 as a Switching Technology in Third-Generation Mobile Access Networks,” *IEEE Communications Magazine*, vol. 37, no. 6, pp. 112–122, June 1999.

[43] S. Nananukul, Y. Guo, M. Holma, and S. Kekki, “Some Issues in Performance and Design of the ATM/AAL2 Transport in the UTRAN,” in *IEEE Wireless Communications and Networking Conference (WCNC2000)*, vol. 2, Chicago, IL, USA, Sept. 2000, pp. 736–741.

[44] MWIF, “IP in the RAN as a Transport Option in 3rd Generation Mobile Systems,” Mobile Wireless Internet Forum (MWIF),” Technical Report MTR-006, Release v2.0.0, June 2001. [Online]. Available: http://www.mwif.org

[45] S. Chen and K. Nahrstedt, “An Overview of Quality of Service Routing for Next-Generation High-Speed Networks: Problems and Solutions,” *IEEE Network*, vol. 40, no. 5, pp. 64–79, Nov. 1998.

[46] G. Apostolopoulos et al., “QoS Routing Mechanisms and OSPF Extensions,” RFC 2676, Aug. 1999.

[47] B. Jamoussi (Editor), “Constraint-Based LSP Setup using LDP,” RFC 3212, Jan. 2002.

[48] D. Awduche et al., “RSVP-TE: Extensions to RSVP for LSP Tunnels,” RFC 3209, Dec. 2002.

[49] “The Network Simulator ns-2 notes, documentation and source codes.” [Online]. Available: http://www.isi.edu/nsnam/ns

[50] “MPLS Network Simulator (MNS),” original by Gaeil Ahn, with MNS v1.0 and MNS v2.0. [Online]. Available: http://flower.ce.cnu.ac.kr/ fog1/mns

[51] “Christian Glomb’s adaptation of MNS v2.0 for ns-2.1b9.” [Online]. Available: http://w4.siemens.de/ct/en/technologies/ic/mpls

[52] “MobiWan: NS-2 extensions to study mobility in Wide-Area IPv6 Networks.” [Online]. Available: http://www.inrialpes.fr/planete/mobiwan

[53] “QoSR (QoS Routing) in ns-2.” [Online]. Available: http://www.tct.hut.fi/tutkimus/ironet/ns2/ns2.html

[54] S. Deering and R. Hinden, “Internet Protocol, Version 6 (IPv6) Specification,” RFC 2460, Dec. 1998.

[55] X. P. Costa and H. Hartenstein, “A simulation study on the peformance of Mobile IPv6 in a WLAN-based cellular network,” *Computer Networks*, vol. 40, no. 1, pp. 191–204, Sept. 2002.